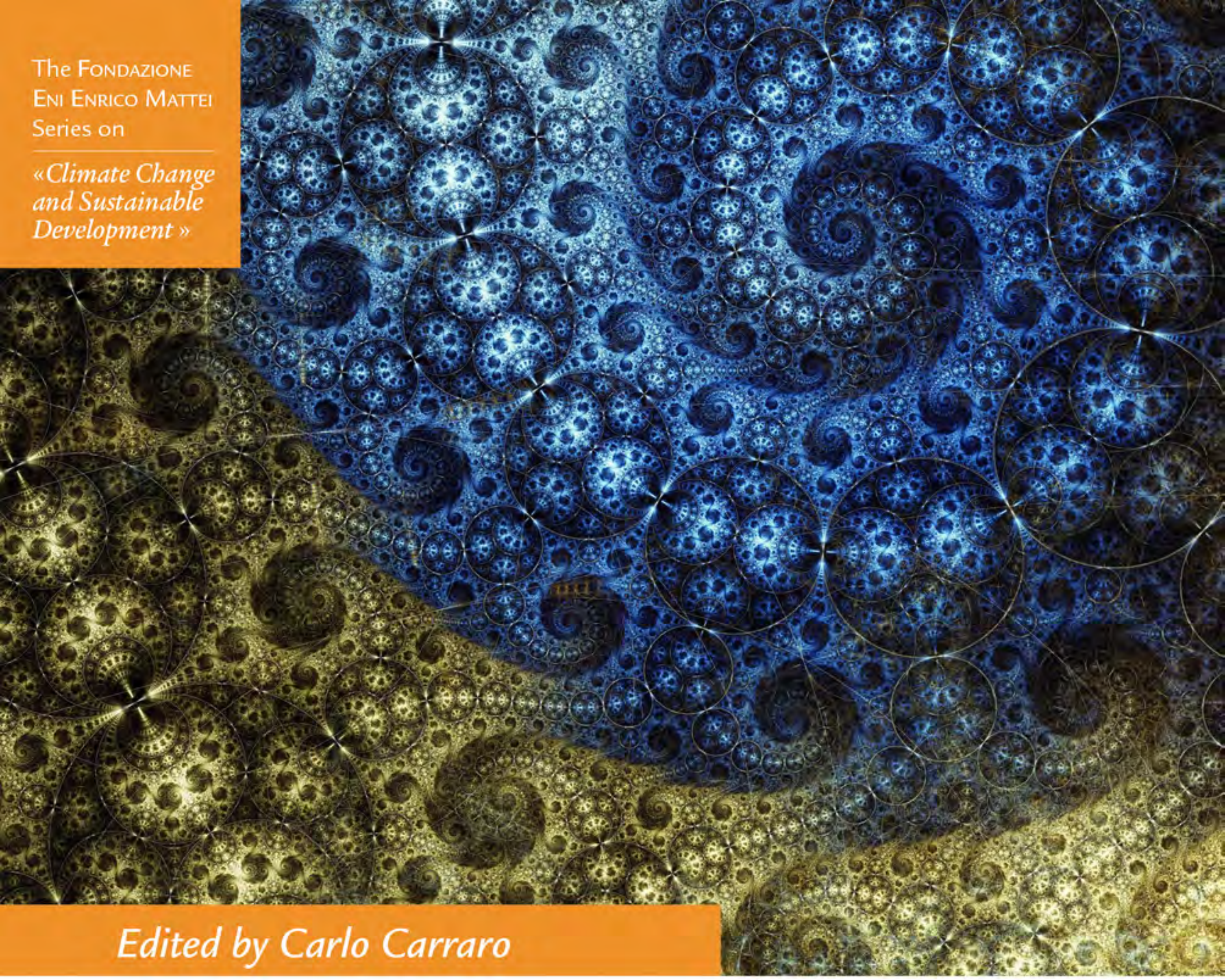


The Fondazione
Eni Enrico Mattei
Series on

*«Climate Change
and Sustainable
Development»*

Edited by Carlo Carraro

Coalitions and Networks

12 papers from 20 years of CTN workshops

FEEM
PRESS

ENGLISH

The **Fondazione Eni Enrico Mattei (FEEM)** Series on

«Climate Change and Sustainable Development»

FONDAZIONE ENI
ENRICO MATTEI

Foreword

Economic development is an essential component of the life of human societies: it is crucial to provide better living conditions to present to future generations. Poverty eradication, better nutrition, access to energy, health and education are objectives almost unanimously shared by all societies, but require economic development to be achieved. Resources in our planet are however finite, and living conditions do not depend only on economic development but also on the quality of the environment. Among the threats to economic development and quality of life in this planet, climate change is certainly the most important one. But other threats, from biodiversity losses to lack of water availability, often closely interrelated with climate change, cannot be neglected when assessing the future prospect of human life on earth. This series' goal is to provide a comprehensive and multidisciplinary approach to sustainable development and to analyze its economic, social and environmental components. FEEM's « Climate Change and Sustainable Development » Series aims indeed at disseminating research carried out and knowledge developed within FEEM's Climate Change and Sustainable Development program. Volumes will move from methodological tools (game theory, integrated assessment models, risk assessment tools, etc.) to economic and policy analysis of measures designed to control climate change, to offset its impacts and, more widely, to support and operationalize sustainable development.

Premessa

Lo sviluppo economico è una componente essenziale della vita delle nostre società: è infatti decisivo per fornire migliori condizioni di vita alle generazioni presenti e future. Ridurre la povertà, migliorare i livelli di nutrizione, fornire a tutti accesso all'energia, all'istruzione e ai servizi sanitari sono obiettivi unanimemente condivisi, ma che richiedono per essere raggiunti uno sviluppo economico diffuso e inclusivo. Le risorse del pianeta sono tuttavia limitate e le condizioni di vita non dipendono solo dallo sviluppo economico, ma anche dalla qualità dell'ambiente in cui viviamo. Il cambiamento climatico rappresenta oggi la più importante tra le minacce allo sviluppo economico e al miglioramento della qualità della vita sulla terra. Ma molte altre variabili, dalla perdita di biodiversità alla mancanza di risorse idriche, spesso dipendenti dal cambiamento climatico, vanno tenute in considerazione se si vuole capire quale possa essere il nostro futuro. L'obiettivo di questa collana è quello di fornire un approccio olistico e multidisciplinare allo sviluppo sostenibile, per poterlo analizzare in tutte le sue componenti: economiche, sociali, ambientali. La collana « *Cambiamento Climatico e Sviluppo Sostenibile* » *è infatti uno degli strumenti con cui la Fondazione Eni Enrico Mattei vuole diffondere la ricerca e la conoscenza sviluppate dal suo programma "Climate Change and Sustainable Development". I volumi di questa collana spaziano quindi dagli strumenti metodologici necessari per valutare le dinamiche dello sviluppo sostenibile (dalla teoria dei giochi ai modelli integrati dell'economia mondiale) fino all'analisi delle migliori misure di policy concepite per controllare il cambiamento climatico, per limitarne i suoi effetti e più in generale per sostenere e concretizzare uno sviluppo economico sostenibile.*

The **Fondazione Eni Enrico Mattei (FEEM)** is a non-profit, nonpartisan research institution devoted to the study of sustainable development and global governance. FEEM's mission is to improve through research the rigour, credibility and quality of decision making in public and private spheres.

Fondazione Eni Enrico Mattei
Corso Magenta 63, Milano – Italy
Tel. +39 02.520.36934
Fax. +39.02.520.36946
E-mail: letter@feem.it
www.feem.it

The opinions expressed in this publication are those of the author(s) alone.

ISBN 9788890991899

Text elaboration by Martina Gambaro. Proof correction Silvia Bertolin, Martina Gambaro and Barbara Racah.

Printed in Milan in March 2015 by Roberto Cremonesi.Co Srl

The Fondazione Eni Enrico Mattei (FEEM) Series on

«Climate Change and Sustainable Development»

Coalitions and Networks

12 papers from 20 years of CTN workshops

Carlo Carraro (ed.)

ENGLISH

FEEM PRESS

Table of Contents

List of Contributors

The affiliations indicated are those effective at the time of publication of this volume.

Scott **Barrett**, The Earth Institute, Columbia University, USA

Francis **Bloch**, Université Paris I and Paris School of Economics, France

Thierry **Bréchet**, Université Catholique de Louvain, Belgium

Carlo **Carraro**, Fondazione Eni Enrico Mattei and Department of Economics, Ca' Foscari University of Venice, Italy

Kfir **Eliaz**, The Eitan Berglas School of Economics, Tel Aviv University, Israel

Taiji **Furusawa**, Graduate School of Economics, Hitotsubashi University, Japan

Andrea **Galeotti**, Department of Economics, University of Essex, UK

François **Gerard**, Department of Economics, Columbia University, USA

Sanjeev **Goyal**, Faculty of Economics, University of Cambridge, UK

P. Jean-Jacques **Herings**, Department of Economics, Universiteit Maastricht, The Netherlands

Matthew O. **Jackson**, Department of Economics, Stanford University, USA

Hideo **Konishi**, Department of Economics, Boston College, USA

Michel **Le Breton**, Gremaq and Idei, University of Toulouse 1, France

Ana **Mauleon**, Université Saint-Louis, and Center for Operations Research & Econometrics, Université Catholique de Louvain, Belgium

Frank H. **Page Jr.**, Department of Economics, Indiana University Bloomington, USA

David **Pérez-Castrillo**, Departament d'Economia i d'Història Econòmica and MOVE, Universitat Autònoma de Barcelona, Spain

Debraj **Ray**, New York University, USA

Ronny **Razin**, Department of Economics, London School of Economics, UK

Domenico **Siniscalco**, Morgan Stanley International and Fondazione Eni Enrico Mattei, Italy

Henry **Tulkens**, Center for Operations Research & Econometrics, Université Catholique de Louvain, Belgium

Vincent **Vannetelbosch**, Center for Operations Research & Econometrics, Université Catholique de Louvain, Belgium

Fernando **Vega-Redondo**, Department of Economics, European University Institute and Bocconi University, Italy

Shlomo **Weber**, Department of Economics, Southern Methodist University, USA and New Economic School, Russia

David **Wettstein**, Department of Economics, Ben-Gurion University of the Negev, Israel

Myrna **Wooders**, Department of Economics, Vanderbilt University, USA

Leeat **Yariv**, Division of Humanities and Social Sciences, California Institute of Technology, USA

Preface

Carlo Carraro, Co-founder, Coalition Theory Network

A beautiful dozen. Twelve papers presented in 20 years of meetings of the Coalition Theory Network (CTN). Twelve seminal contributions to the theory of coalitions and networks. Not necessarily the twelve best papers. Certainly many other excellent papers have been presented and discussed in these 20 CTN workshops. Nevertheless, these papers well represent the story of the Coalition Theory Network, from the origins to the various evolutions in these 20 years. In addition, they help to understand the achievements of the Coalition Theory Network and highlight the importance of the various CTN partners.

But let's start from the beginning. Twenty years ago! The Coalition Theory Network was indeed founded in 1995, when FEEM joined CORE – University of Louvain-la-Neuve in organizing a workshop on coalition formation and environmental games focused on the analysis of international environmental agreements and climate negotiations. The success of the workshop induced the organisers to widen the focus of the following CTN meetings to the burgeoning applications of coalition and network theory, and to undertake the formal creation of the Coalition Theory Network. The yearly meetings have continued for 20 years, hosted in turn by the partner institutions, among which those that have joined in the meantime: after FEEM and CORE in 1995, GREQAM – University of Aix-Marseilles and CES – University Paris I joined in 1999, MOVE – Universitat Autònoma de Barcelona in 2000, Maastricht University and Vanderbilt University in 2006, and CSDSI – New Economic School of Moscow joined CTN in 2014.

Year after year, CTN has progressively become a reference point for the study of network and coalition formation and their applications. Its workshops have been attended by an increasing number of scholars. Participation in CTN workshops has become highly selective, with less than 30% of submitted papers accepted for presentation. Thanks to its success, CTN is now a well established association of eight high-level scientific and academic institutions, whose aim is the advancement and dissemination of research in the area of network and coalition theory.

All of the most important contributors to the theory of coalitions and networks have attended one or more CTN workshops. Most have delivered keynote lectures, all have presented papers that have been later published in top economic journals. The CTN web site (http://www.coalitiontheory.net/) is certainly the best specialised repository of knowledge on coalitions and networks.

How, then, could 12 papers be selected out of more than 500 papers presented at the CTN workshops? And why? Even though quite a small drop in the CTN ocean, this sample of papers is very informative about the story and the accomplishments of the Coalition Theory Network. Most of the twelve papers belong to the first years of CTN: they are a sort of memory of the foundations of this network. Foundations lying on the support received from the network partners above all. But also lying on the contribution of specific persons. From Henry Tulkens (CORE) to Domenico Siniscalco (FEEM), from Francis Bloch (GREQAM and now Paris I) and Antoine Soubeyran (GREQAM) to Salvador Barberà (Barcelona), from Scott Barrett (now at Columbia University) to Shlomo Weber (now at the New School of Economics in Moscow), from Myrna Wooders (Vanderbilt) to Hubert Kempf (Paris I), from Debraj Ray (NYU) to Matt Jackson (Stanford) and Michael Finus (now at University of Bath). And of course many others, whose names can be found in this volume or in the program of the CTN workshops available at http://www.coalitiontheory.net/.

The twelve papers, therefore, have been certainly chosen for their quality, but not only that. They first and foremost represent the eight partners of CTN. They outline the ideas and the vision at the origin of the network. They highlight the research topics that drove the interest on coalitions first and on networks then.

For example, the prominence given to contributions to the theory of international environmental coalitions can be explained because this was the seed that gave rise to many other ideas and theoretical developments. International environmental cooperation, as well as international cooperation over other global economic, social or military issues, was, and still is, increasingly important worldwide. The range of topics on which negotiations to achieve a substantial degree of cooperation among countries and regions are underway is indeed wide. Transnational issues, such as trade and financial flows liberalization, migration, technological co-operation, development aid, disease control and climate change are the most important issues discussed in G-8, G-20 and other international meetings.

The common feature of these issues is a high degree of interdependence among countries: in general, the welfare of each country depends on its own action as well as on the action of any other country. As a consequence, in most cases, unilateral policies can be jeopardized and possibly made useless by the other countries' reaction. This is the well-known "tragedy of the commons". International cooperation, which makes policy more effective and can also redistribute the resulting gains among the cooperating countries, is therefore welfare improving. How to achieve these welfare improvements is therefore a relevant research question, which drove the first attempts by CTN researchers to identify mechanisms and policies to foster international cooperation.

The study of countries' interactions when dealing with an international or global economic problem can obviously be represented as a game, and the emergence of cooperation as the decision to form a coalition. That's why it became important to study the formation of coalitions. But the study of agents' behaviour when forming a coalition led CTN researchers to analyse other forms, simpler or more complex, of agents' interaction. In a coalition, all players forming a coalition interact with all other coalition members and, as a group, with all the other players of the game. In a network, some players cooperate with others, other players possibly with only one player, some players may not even interact with others, while still belonging to the same game. These more detailed and articulated forms of interactions are studied by the theory of networks, whose applications have been wide, possibly even wider than those of coalition theory.

This is why in this volume you will find chapters focusing on trade networks, or on the theory of organisations, or on homophily and friendship. The theory of networks has indeed a large spectrum of applications, and the properties of networks have been the subject of many contributions to CTN workshops (partly captured in this book as well). Network theory aims to provide a unified framework for analyzing the relation between agents' position in the network and their actions and welfare. More generally, a model is needed to explain how the whole structure of the network (or the beliefs that agents hold in this structure) affect agents' behavior and welfare. The study of network formation and of the games played in networks under local and limited information is indeed one of the most challenging and frequently studied issues at present. Applications of the theory of networks also include the governance of economic unions, the formation of industrial cartels and collaborations, the patterns of racial integration in social networks, and the endogenous evolution and structure of institutions, etc.

This volume therefore represents a tribute to research developed by the Coalition Theory Network, and presented at the CTN annual workshops, on the occasion of the 20th anniversary of its foundation. It is a tribute to the eight partners of the Coalition Theory Network and to all the colleagues who have contributed to its success. It is finally a tribute to the wide array of useful and interesting applications of the theory of coalitions and networks, partly underutilised by applied economists, that CTN helped to develop and disseminate.

The success of the Coalition Theory Network is certainly explained by the vision and commitment of the eight CTN partners, and by the intellectual achievements of the many game theorists and economists who attended CTN workshops. However, CTN is also strongly indebted to the work and dedication of Silvia Bertolin, who has been in charge of the CTN secretariat for about fifteen years and who has managed its web site, prepared the quarterly newsletter and, above all, maintained friendly and cooperative links among all partners. And even this book would not have been published without Silvia's work, perfectly complemented by the excellent contributions of Martina Gambaro and Barbara Racah.

On its 20th anniversary, the Coalition Theory Network is still very lively and increasingly attractive for young researchers. For its 20th anniversary workshop, the number of submissions has achieved a record number. And partnership is likely to further develop. I hope this volume will help retrace the past of the Coalition Theory Network, while at the same time stimulating its future developments and success.

Venice, March 2nd 2015

Strategies for the International Protection of the Environment

Carlo Carraro and Domenico Siniscalco

Paper presented at the CTN Workshop 1994. Reprinted from *Journal of Public Economics* **52**(3), October 1993, 309–328, with permission from Elsevier.

This paper analyses profitability and stability of international agreements to protect the environment in the presence of trans-frontier or global pollution. Each country decides whether or- not to coordinate its strategy with other countries. A coalition is formed when conditions of profitability and stability (no free-riding) are satisfied. It is shown that such coalitions exist; that they tend to involve a fraction of negotiating countries; and that the number of signatory countries can be increased by means of self-financed transfers. However, expanding coalitions requires some form of commitment. Such schemes of commitment and transfers can even lead to cooperation by all countries.

1. Introduction

A large amount of pollutants are discharged in the atmosphere and water systems, as a result of human activity in each country. The emissions often affect other countries, as well as the global environment. In economic terms, each polluting country benefits from using the environment as a receptacle for emissions but, at the same time, is also damaged by environmental deterioration. While the benefit is related to domestic emissions only, the damage is related to both domestic and foreign emissions which negatively affect the environment. Hence, a problem of international externalities arises which, in the present

We are grateful to K.J. Arrow, C. D'Aspremont, M. Hoel, K.G. Maler, P. Michel and two anonymous referees, for helpful comments and suggestions on a previous version of this paper. The usual disclaimers apply.

institutional setting, can be solved only by voluntary agreements among sovereign countries. Such agreements have been quite common in recent years,[1] and they seem to share some features: they are often characterized by cooperative behaviour among the individual countries involved; they usually have only a sub-group of the negotiating countries as signatories (partial cooperation); and they tend to use various forms of transfers, typically to the developing countries, as a key instrument for increasing the number of signatories.

The existing literature on the protection of the international environment does not sufficiently convey the characteristics of international agreements mentioned above. The traditional contributions on trans-national commons describe countries' environmental interaction as a one-shot Prisoner's Dilemma, where free-riding inevitably leads to the 'tragedy of commons' (for a discussion, see Ostrom, 1990). In more recent works on the subject, the repetition of the Prisoner's Dilemma, under appropriate assumptions, can enlarge the set of equilibrium outcomes, and characterize situations where all countries cooperate (for a discussion, see Maler, 1989; Barrett, 1992). Partial cooperation and the role of transfers, however, are not usually considered in either approach.

This paper presents a general framework to analyse the profitability and stability of international agreements to protect the environment in the presence of trans-frontier or global pollution. In our analysis, international negotiations are modelled as games in which sovereign countries bargain over emission control. With respect to emission control, countries may choose to act either cooperatively or non-cooperatively. In the former case, cooperative agreements can involve a sub-group of countries, whose number can be expanded by means of 'self-financed' welfare transfers. The framework re-interprets results which were already presented in the recent environmental literature with reference to specific cases (e.g. Maler, 1989; Newbery, 1991; Barrett, 1991, 1992; Hoel, 1992; Kaitala et al., 1992). More importantly, it provides new results on the emergence of partial cooperation, transfers, and the role of commitment.

The main conclusions of the paper can be summarized as follows:

(i) the strategic interaction among countries in a common environment does not necessarily lead to the 'tragedy of commons', but there is a wide range of possible voluntary agreements to control emissions;

(ii) beyond non-cooperative emission control, there exist partial cooperative agreements among sub-groups of countries (coalitions) which are profitable and stable;

(iii) gains from partial cooperation can be used to expand existing coalitions by

[1] Experts quote more than 150 international agreements which have been signed to protect the environment in various regions. The protocols on CFC, and the ongoing negotiation on global warming provide examples of agreements at the global level.

> inducing other countries to cooperate using self-financed welfare transfers. To sustain broader coalitions by means of transfers, however, a minimum degree of commitment must be introduced into the game, thus changing its rules. The various forms of commitment proposed in the paper are less demanding than a commitment of cooperation by all players, since they may involve only a fraction of the cooperating countries. Such schemes of partial commitment and transfers can even lead to cooperation by all countries.

The paper is organized as follows. In Section 2, the general framework is introduced and some kinds of agreements which can lead to pollution control are defined. Results on· stable coalitions, the role of transfers and commitment are also provided. In Section 3, the main results and implications of the proposed framework are discussed. Finally, in Section 4, some extensions of the model are proposed, together with the scope for further work.

2. The Analytical Framework

2.1 Players, payoffs, strategies

Consider n countries ($n \geqq 2$) that interact in a common environment, and bargain over emission control of a specific pollutant. Each country i benefits from using the environment as a factor of production and as a receptable for emissions. The welfare of each country, however, is negatively affected both by its own emissions x_i and by other countries' emissions x_{-i}, where x_{-i} is the vector $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Country i's benefit and damage can be represented in the welfare function $P_i(x) = B_i(x_i) - D_i(x_i, x_{-i})$, where $B_i(x_i)$ denotes benefits arising from the use of the environment for production and consumption activities; $D_i(x_i, x_{-i})$ denotes damages (welfare losses) resulting from pollution emissions;[2] and $x \equiv (x_1 \ldots x_n) = (x_i, x_{-i})$ is the vector of all countries' emissions.

Let us consider the benefit function $B_i(x_i)$. It is assumed that a reduction in pollution, which can be achieved through domestic environmental policies, is costly and reduces benefits. The benefit function, which depends on abatement costs, is related to the level of technology, the economic structure, the level of development, and the endowment of resources of a country. By technology we mean more than the mechanical process of turning inputs into outputs; we mean useful knowledge and experience, institutions and organizational structures, norms and values that govern the processes of production and exchange.

The damage function $D_i(x_i, x_{-i})$ depends on a country's perception of the effects of emissions of a given pollutant, as well as on the evaluation of such

[2] The parameters of the damage function may include the i-th row of the transportation matrix $A = \{\alpha_{ij}\}$, where α_{ij} is the share of the country j's emissions affecting country i.

effects. Consequently, the damage function is based on a subjective evaluation of environmental goods. The specific functional form of $D_i(\cdot)$ can be determined using appropriate models of environmental impact evaluation. In such models, an index usually summarizes both the measurement of the relevant physical damages and their evaluation.

Countries consider one pollutant at a time. Let δ_i, be the maximum level of pollution emissions for country *i*. It is computed by maximizing environmental benefits $B_i(x_i)$ without taking into account the associated costs evaluated through the function $D_i(\cdot)$. Moreover, δ_i can be seen as a measure of a country's size and level of development. The 'emission game' between the n countries is defined by a triple (N, S, P), and by appropriate rules: as usual, $N = \{1 \ldots n\}$ is the set of players, $S = S_1 \times \cdots \times S_n$, where $S_1 = [0,\delta_i]$, is the strategy space, and $P = (P_1(x) \ldots P_n(x))$ is the payoff vector. Complete information is assumed. Problems arising from asymmetric information will briefly be mentioned in Section 4.

By taking into account reciprocal externalities, a country may decide whether or not to cooperate with other countries in order to reduce total emissions. At this stage, we assume that cooperative agreements are not binding. The decision regarding whether or not to. cooperate is the outcome of a 'metagame' in which each country anticipates the choice (cooperative or non-cooperative) of the other countries, and the relative outcomes in terms of emission levels.[3]

Let us begin by analysing the outcomes of the game under alternative strategic combinations. First, we assume that countries play simultaneously and non-cooperatively. Thus, country *i*'s optimal level of emissions is determined by equating marginal benefits with marginal costs, given the emission levels set by the other countries. The solution of the system of first order conditions determines the Nash equilibria of the game. For simplicity's sake, we assume the equilibrium to be unique. The Nash equilibrium of the non-cooperative game can also be determined by computing the fixed point of countries' best-reply functions. Let $R_i(x)$, $x = (x_1 \ldots x_n)$, the country *i*'s best-reply function, where $R_i(x) = \{x_1 \colon P_i(x_i, x_{-i}) \geqq P_i(s_i, x_{-i})$, for all $s_i \in S_i\}$. The non-cooperative equilibrium x^0 is defined by $x^0 = \mathrm{R}(x^0)$, where $R(x) = (R_1(x) \ldots R_n(x))$.

Alternatively, countries can decide to set emissions cooperatively. In this case, we assume that a bargaining process takes place in order to achieve a Pareto optimal outcome. The bargaining process may lead to the formation of a

[3] We restrict our analysis to one-shot games. In terms of additional equilibrium outcomes, analysing repeated games would be fruitful only if appropriate trigger or stick/carrot strategies could sustain cooperation as an equilibrium outcome. The level of emissions, however, can hardly be conceived as a trigger variable which can be increased strategically in response to other countries' defection. Some reasons are the following: firstly, emission reduction, in cases such as $C0_2$ or CFC, involves substantial and irreversible investments. Secondly, expanding emissions as a retaliation could generate environmental damage primarily to the triggering country. Finally, an increase in emissions can hardly be used as a selective punishment. Other effective punishments (e.g. trade protectionism) could be even more costly for the triggering country and therefore, not credible. For these reasons, we believe that trigger or stick/carrot strategies are not particularly helpful in sustaining environmental cooperation.

coalition among *j* countries, where *j* goes from 2 (the smallest feasible coalition) to *n* (when all countries set emissions by taking into account reciprocal externalities). We define 'full cooperation' as a situation in which a coalition formed by n countries emerges, while a 'partial cooperation' is a coalition formed by $2 \leqq j < n$ countries. In this work, we determine the cooperative outcome of the game by using the Nash bargaining solution.[4] Moreover, we use the non-cooperative equilibrium $x^0 = (x_1^0 \ldots x_n^0)$ as the threat point of the bargaining process.[5] Stated more formally, *j* countries will act cooperatively when they set emissions maximizing the joint product of the difference between $P_j(x)$ and P_i^0, where P_i^0 denotes the non cooperative welfare.

Before setting emission levels, each country must therefore decide whether or not to act cooperatively. This decision can be modelled by defining a 'metagame' in which countries choose between the cooperative and the non cooperative strategy by anticipating the outcomes of the related emission game.[6]

2.2 Profitability and stability of coalitions

Let $P_i(j)$ be country *i*'s welfare when it decides to cooperate, and $j - 1$ when other countries also cooperate; whereas $Q_i(j)$ is its welfare when country *i* does not join the coalition formed by *j* countries. Moreover, let *j* stand for the set of cooperating countries, and J^0 denote the set of countries that play non-cooperatively. Let us suppose, for simplicity, that all countries are symmetric, i.e. the welfare function is not country specific. We do not therefore index the welfare functions *P* and *Q* and their parameters.

A minimum requirement must be met for an environmental coalition to be formed: the welfare of each country signing the cooperative agreement must be larger than the non-cooperative welfare. In other words, country *i* gains from joining the coalition, with respect to its position when no countries cooperate, if $P(j) > P^0$. This leads to:

Definition 1. A coalition formed by *j* players is ***profitable*** if $P(j) > P^0$ for all countries belonging to *J*.

Of course, this only represents a minimum requirement that may not suffice to induce many countries to sign a cooperative agreement. The main problem preventing

[4] This assumption is not crucial, because, in the rest of the analysis, all countries will be assumed to have the same benefit and damage functions. As was pointed out by a referee, in this case any bargaining solution in the literature would give the same result.

[5] In a two-player game, this means interpreting Rubinstein's alternating offers model as a model in which players face a risk that, if the agreement is delayed, then the opportunity they hope to exploit it jointly may be lost (see Binmore et al. ,1986).

[6] Most environmental studies model this 'metagame' as a one-shot Prisoner's dilemma, in which non-cooperation is the dominant strategy.

the formation of any coalition is the possibility of free-riding by some countries. Free-riding can be explained as follows: since one country can profit from the reduction of emissions by cooperating countries, it has an incentive to let other countries to sign the cooperative agreement. If all countries are symmetric, no cooperation takes place. In other words, the 'metagame' in which countries choose between cooperation and non-cooperation is represented as a Prisoner's dilemma. As we will see, however, this representation of countries' strategic choice may not be appropriate.

The problem can also be stated more formally. For each country, the crucial comparison is between $P(j)$, the country's payoff for belonging to the j-coalition, and $Q(j-1)$, the country's payoff when it exists the coalition, and lets other $j-1$ countries sign the cooperative agreement. Let $Q(j-1)-P(j)$ be a country's incentive to defect from a coalition formed by j players, whereas $P(j+1)-Q(j)$ is the incentive for a non-cooperating country to join a j-coalition [which, consequently, becomes a $(j+1)$-coalition]. Thus, a stable coalition can be defined as follows:

Definition 2. A coalition formed by j players is ***stable*** if there is no incentive to defect, i.e. $Q(j-1)-P(j)<0$, for all countries belonging to J, and there is no incentive to broaden the coalition, i.e, $P(j+1)-Q(j)<0$, for all countries belonging to J^0.

This definition corresponds to that of cartel stability presented in the oligopoly literature (see D'Aspremont and Gabszewicz, 1986, a similar definition is also used in Barrett, 1991).

It has been shown that under fairly general conditions stable coalitions exist (see Donsimoni et al., 1986). However, this does not satisfactorily address the problem of protecting international commons because, as has been demonstrated both in the oligopoly and in the environmental literature (see, for example, D'Aspremont et al., 1983; D'Aspremont and Gabszewicz, 1986; Carraro and Siniscalco, 1991; Hoel, 1992), stable coalitions are generally formed by $j \leqq n$ players, where j is a small number, regardless of n.[7] This leads us to the following question: can the j cooperating countries expand the coalition through self-financed welfare transfers to the remaining players?

Given the previously stated rules of the game, the answer is no. This is demonstrated by the following proposition:

Proposition 1. *Suppose no countries can commit to the cooperative strategy. Then, in this case, no self-financed transfer T from the j cooperating countries to the other non-cooperating countries can successfully enlarge the original coalition.*

[7] More satisfactory results are presented in Barrett (1991), who shows that, under certain conditions, and given a specific functional form for the welfare function, large stable coalitions exist.

Proof. For the transfer to be self-financed, it cannot be larger than the gain that the j players obtain from moving to a $(j+1)$-coalition. Furthermore, in order to add one player to a j-coalition, the transfer T must be larger than the loss incurred by the $j+1$ player by entering it. These two conditions yield:

$$[P(j+1) - P(j)] \geqq T > Q(j) - P(j+1) \qquad (1)$$

This condition makes it possible to self-finance an enlarged coalition. However, is this broadened coalition stable? The $j+1$ player does not defect if the transfer is larger than $Q(j) - P(j+1)$. However, by definition of stable coalition, $P(j+1) < Q(j)$: the j players of the original coalition have therefore an incentive to defect; their incentive being greater because of the transfer made to the $j+1$ player. Hence, the $(j+1)$-coalition is unstable. ■

This leads to the following conclusion: welfare transfers from countries belonging to a stable coalition to non-cooperating countries cannot be used to expand the initial coalition, unless the rules of the game are changed. There are various rules that can lead to the formation of larger stable coalitions. We will focus our attention on the role of commitment. If all countries were committed to cooperation, obviously no free-riding would exist. We show, however, that there are several, less demanding forms of commitment which, if associated with appropriate welfare transfers, can lead to large stable coalitions. There are two crucial elements in our analysis: partial commitments (only a subset of the n countries commit to co-operation), and welfare transfers (bribing).

2.3 Expanding coalitions

Hereafter, we shall analyse the implications of four types of commitment, that could serve as possible blueprints for environmental cooperation. Of course, other types of institutional mechanisms could be proposed as well.

The four types of commitments are:

(i) Only the j countries belonging to the stable coalition commit to cooperation (stable coalition commitment).

(ii) The j countries are committed to cooperation and any new signatory, as soon as it enters the expanded coalition, must commit to cooperation as well (sequential commitment).

(iii) The number of committed countries is such that appropriate transfers can induce all other countries to cooperate (full-cooperation minimum commitment).

(iv) A subset of non-cooperating countries commits to transfer welfare in order to induce the remaining non-signatories to cooperate, and to guarantee the stability of the resulting coalition (external commitment).

As has already been shown, it is necessary to impose constraints on the amount of transfers allowed: if no restriction were imposed, all non-signatories could be bribed. Therefore, we assume:

(i) transfers must be self-financed, i.e. the total transfer T must be lower than the gain that the committed countries obtain from expanding the coalition;

(ii) the move to a larger coalition must be Pareto-improving, i.e. all countries must be better off than in the situation preceeding the coalition expansion, and better off with respect to the case of no-cooperation (the larger coalition must be profitable); and

(iii) committed countries choose the transfer in order to maximize the number of signatories (given the above two constraints).[8]

Given these restrictions, the following question can be asked: Under what conditions can partial commitments and transfers expand the initial j-coalition? Some answers are provided by the following propositions:

Proposition 2 – Stable coalition commitment. *Suppose the j countries belonging to the stable coalition are committed to cooperation. If $P(j + s) > P(j)$ and $Q(j + s) > Q(j)$ for all positive s, $s \leqq n - j - 1$, then at most r countries can be induced to join the initial coalition, where r is the largest integer satisfying:*

$$r < j[P(j + r) - P(j)]/[Q(j + r - 1) - P(j + r)] \tag{2}$$

Proof. The initial j countries can use their gains resulting from broadening the coalition in order to finance other countries' cooperation. This gain is equal to $j[P(j + r) - P(j)]$, and it is positive if $P(j + r) > P(j)$. In order for transfers to be self-financed, they must be greater than the incentive to defect for the r countries that have to enter the coalition. This incentive is equal to $r[Q(j + r - 1) - P(j + r)]$. Hence:

$$j[P(j + r) - P(j)] > r[Q(j + r - 1) - P(j + r)] \tag{2'}$$

Moreover, the maximum transfer $j[P(j + r) - P(j)]$ must be larger than the total loss suffered by entering countries; this loss is equal to $r[Q(j) - P(j + r)]$.

This condition can be expressed as follows:

[8] Notice that transfers enter a country's payoff in an additive way. Otherwise, each cooperating country would maximize its payoff with respect to both emissions and transfers. This would be the case if transfers were carried out using policy instruments which interact with other economic variables, and with emissions in particular. Such an extension of the paper is discussed in the concluding section.

$$j[P(j + r) - P(j)] > r[Q(j) - Q(j + r - 1) + (Q(j + r - 1) - P(j + r))] \tag{3}$$

Notice that $[Q(j) - Q(j + r - 1)] \leq 0$ and $[Q(j + r - 1) - P(j + r)] > 0$. Hence, (2') implies (3). Since (2) implies (2'), the proposition is proved. The newly entered countries have no incentive to defect, and benefit from joining the coalition; moreover, the initial cooperating countries benefit from expanding the coalition, and are committed to cooperation. Therefore, the new equilibrium constitutes a Pareto improvement. ■

The meaning of Proposition 2 is the following: starting from a *j*-coalition, the commitment of its members induces, through appropriate transfers, other *r* countries to join the initial coalition; notice that *r* is larger, the larger the gain attained from expanding the coalition, and the lower the incentive to defect from it.

Proposition 3 – Sequential commitment*. Suppose that a stable coalition formed by j countries exists. If each country, when entering the coalition, commits to cooperation, and if* $P(j + s) > P(j + s - 1)$ *for all* $1 \leqq s \leqq r$*, then at most r countries can be induced to join the initial coalition if*

$$j + s > [Q(j + s - 1) - P(j + s - 1)]/[(j + s) - P(j + s -1)] \tag{4}$$

for all $1 \leqq s \leqq r$*, and, when* $j + r < n$*,*

$$j + r < [Q(j + r) - P(j + r + 1)]/[P(j + r + 1) - P(j + r)] \tag{4'}$$

Proof. Suppose that *j* countries form a stable coalition, and are committed to cooperation. For one more country to be induced to cooperate, the transfer must be lower than or equal to the total gain derived from moving to a $(j + 1)$-coalition, and larger than the loss incurred by the entering country. This gives

$$j[P(j + 1) - P(j)] > Q(j) - P(j + 1) > 0 \tag{5}$$

Suppose this condition is satisfied and that the $(j + 1)$th country, when entering the coalition, commits to cooperation. An additional country can be induced to join the coalition if

$$(j + 1)[P(j + 2) - P(j + 1)] > Q(j + 1) - \mathrm{P}(j + 2) > 0. \tag{6}$$

By iterating this reasoning, one can say that *r* countries can be induced to join the initial coalition through appropriate transfers, and the sequential commitment, if

$$(j + s - 1)[P(j + s) - P(j + s - 1)] > Q(j + s - 1) - P(j + s) > 0 \tag{7}$$

for all $1 \leqq s \leqq r$. If this condition is satisfied for all s such that $j < s \leqq n - j$, then all countries will join the coalition. Thus, full cooperation can be achieved.

Otherwise, the coalition cannot be expanded further if

$$(j + r)[P(j + r + 1) - P(j + r)] < Q(j + r) - P(j + r + 1) \tag{7'}$$

Full cooperation cannot always be achieved because there may exist values of r for which the gain from further broadening the coalition is lower than the loss incurred by the entering country, i.e. the transfer would not be sufficient to induce one more country to join the coalition.

Eq. (7) can be re-written as:

$$j + s > [Q(j + s - 1) - P(j + s - 1)]/[P(j + s) - P(j + s - 1)] \tag{8}$$

for all $1 \leqq s \leqq r$, and

$$j + r < [Q(j + r) - P(j + r + 1)]/[P(j + r + 1) + P(j + r)]. \tag{8'}$$

It should be noted that, when the $(j + r)$-coalition is formed, all countries gain from broadening the coalition. Moreover, given the assumption regarding commitment, the coalition is stable. ■

Proposition 3 has two implications. First, as in the case of stable coalition commitment, the expanded coalition is larger, the larger the gain from moving to a wider coalition, and the lower the incentive to defect from it. Secondly, this form of commitment is more demanding than the previous one, since it eliminates the problem of guaranteeing the coalition stability (by the sequential commitment assumption). Hence, unless (4') holds for some $r < n - j$, full cooperation can be achieved.

Proposition 4 – Full-cooperation minimum commitment. *If $Q(n - 1) > Q(i)$ for all positive $i < n - 1$, at least a fraction*

$$i/n > [Q(n - 1) - P(n)]/[Q(n - 1) - P(i)] \tag{9}$$

of the n countries must be committed to a cooperative strategy for all n countries to cooperate.

Proof. Let us assume that i countries are committed to cooperation regardless of the size of the coalition that is formed. Suppose the i countries offer transfers to induce the other $(n-i)$ countries to join the coalition. The total transfer T must be less than or equal to the gain that the i players achieve from entering the n-coalition. Moreover, this transfer should compensate the $(n-i)$ players for the loss from joining the coalition, and should also offset their incentive to defect from the n-coalition. In order to compensate the $(n-i)$ players for the loss from joining the n-coalition, the following condition must be met:

$$i[P(n) - P(i)] \geqq T > (n-i)(Q(i) - P(n)) \tag{10}$$

This condition ensures that the enlarged coalition is self-financed. It can be re-written as

$$i(Q(i) - P(i)) > n(Q(i) - P(n)) \tag{10'}$$

In order to offset the incentive to deviate from the n-coalition, the gain $P(n)$ plus the transfer $i(P(n) - P(i))/(n-1)$ must be larger than the defector's welfare $Q(n-1)$, i.e. rearranging the equation

$$i(Q(n-1) - P(i)) > n(Q(n-1) - P(n)) \tag{11}$$

which is equivalent to (9). Notice that both sides of the equation are positive. Eq. (11) can be re-written as

$$i[(Q(n-1) - Q(i)) + (Q(i) - P(i))] > n[(Q(n-1) - Q(i)) + (Q(i) - P(n))] \tag{11'}$$

Let us show that (11') implies (10'). Assume that (11') holds as an equality. Thus, we can solve it with respect to $Q(i) - P(i)$. This expression can then be substituted into eq. (10'), yielding $(n-i)[Q(n-1) - Q(i)] > 0$, which is satisfied for all positive $i < n-1$. Hence, condition (9) guarantees that both the financing condition (10) and the no-defection condition (11) are satisfied. As a consequence, the (11') players that join the initial coalition have no incentive to defect. The initial i players are instead committed to cooperation. Finally, the move to the n-coalition is a Pareto improvement. ■

Proposition 4 states that there exists a minimum number of countries that should commit to cooperation for all the remaining countries to be induced to cooperate using appropriate transfers. The minimum number of committed cooperating countries decreases as the gain from moving from the i-coalition to the n-coalition increases, and as the incentive to defect from the latter decreases.

The last case we would like to explore considers how an environmental coalition can be expanded using external commitment. Non-cooperating countries gain when the cooperative agreement is expanded (because they receive less emissions). Consequently, a subset of these countries could find it profitable to induce other non-signatories to enter the coalition, and to secure its stability. Suppose, then, that a fraction of non-cooperating countries commits to finance environmental cooperation (emission reduction in other countries). What is the largest number of countries that can be induced to form a stable coalition?

Proposition 5 – External commitment. *Suppose that j countries cooperate, and that $n - j - r$ non-cooperating countries commit to transfer welfare both to induce r non-signatories to cooperate, and to guarantee the stability of the resulting coalition. If $P(j + s) > P(j)$ and $Q(j + s) > Q(j)$ for all positive $s \leqq n - j - 1$, a stable $(j + r)$-coalition can be formed if*

$$(j + r)/n < 1/(1 + \theta), \qquad (12)$$

where $\theta = [Q(j + r - 1) - P(j + r)]/[Q(j + r) - Q(j)]$.

Proof. Assume there exists a stable j-coalition. The $(n - j - r)$ countries who do not join the coalition gain from financing, through appropriate transfers, a larger coalition, if $Q(j + r)$, their payoff when the $(j + r)$–coalition is formed, less $(Q(j + r - 1) - \mathrm{P}(j + r))(j + r)/(n - j - r)$, the transfer to the $(j + r)$ cooperating countries, is larger than $Q(j)$, their payoff before broadening the coalition. This is true if (12) holds. Moreover, $Q(j + r) - (Q(j + r - 1) - P(j + r))(j + r)/(n - j - r)$ must be larger than $P(j + r + 1)$, i.e. no additional countries want to join the coalition. This is true if

$$(n - j - r)[Q(j + r) - P(j + r + 1)] > (j + r)[Q(j + r - 1) - P(j + r)], \qquad (13)$$

which can be written as

$$n[Q(j + r) - P(j + r + 1)] > (j + r)[Q(j + r) - P(j + r) + Q(j + r - 1)$$
$$-P(j + r + 1)]. \qquad (13')$$

Comparing (12) and (13'), it is easy to see that (12) implies (13') [and therefore (13)].

Finally, we have to prove that the players in the $(j + r)$-coalition have no incentive to defect. This is true if $P(j + r)$, the welfare when the $(j + r)$-coalition is formed, plus $Q(j + r - 1) - P(j + r)$, the transfer each cooperating country receives, is not lower than $Q(j + r - 1)$, the welfare that each cooperating country would receive by defecting from the coalition. This implies $Q(j + r - 1) \geqq Q(j + r - 1)$, which obviously holds (we assume that whenever a country is indifferent between cooperation and defection, it cooperates). Moreover, $P(j + r) + [Q(j + r - 1) - P(j + r)]$ is larger than $P(j)$, the welfare that countries in the stable coalition received before its expansion, because $P(j + r) > P(j)$ by assumption and $Q(j + r - 1) - P(j + r) > 0$ by the stability condition; it is also larger than $Q(j)$, the welfare that countries entering the coalition received before, because $Q(j + r - 1) > Q(j)$ by assumption.

As a consequence, all players in the $(j + r)$-coalition do not defect, all players outside the coalition do not want to join it, and the move to a $(j + r)$-coalition constitutes a Pareto improvement. ■

The conclusion reached is the following: r additional countries can be induced to cooperate, and the $(j + r)$-coalition is stable if the remaining $(n - j - r)$ non-cooperating countries commit to carry on appropriate transfers to the $(j + r)$ cooperating countries. The dimension of the resulting coalition increases as the incentive to defect from a $(j + r)$-coalition decreases, and as the gain that non-cooperators achieve from moving to a $(j + r)$-coalition increases.

3. Results and Applications to Environmental Agreements

3.1 Some general comments

The framework proposed in the previous section aims at explaining the emergence of environmental cooperation without the help of trigger or stick/ carrot mechanisms. Our approach was chosen because, to our knowledge, no international negotiation to protect the environment has ever used pollution as a triggering variable, and because partial cooperation and transfers seem to be common features of many recent agreements in this field.

The crucial steps of the analysis are two: we first characterize stable coalitions, and give some formal conditions for their existence (subsection 3.2). We then use such stable coalitions as a starting point for wider coalitions, expanded by means of transfers and commitments (subsection 3.3). The existence of small stable coalitions is a result already obtained in the most recent environmental literature. Subsection 3.3 on expanded coalitions provides some new results.

The whole framework deserves some general comment, at this stage. If stable coalitions exist, the 'metagame' in which countries decide whether or not to cooperate is not a Prisoner's dilemma. Assume that a j-stable coalition is formed: by definition, no incentive to defect exists ($Q(j - 1) < P(j)$). All countries, however, have an additional incentive not to cooperate. Since non-cooperating countries gain from the others' cooperative behaviour, each country has an incentive to !et other countries form the coalition ($Q(j) > P(j)$). This is not a Prisoner's dilemma because the situation in which one group of countries cooperates and the others do not is an equilibrium of the 'metagame'. This is shown by the 2 x 2 example in Table 1.

Table I

		Country h	
		C	N
Country i	C	$P(j+1), P(j+1)$	$P(j), Q(j)$
	N	$Q(j), P(j)$	$Q(j-1), Q(j-1)$

In this table, C and N denote the cooperative and non-cooperative strategies, respectively, and the payoff pairs indicate countries' welfare, as defined in the previous section. Table 1 represents a situation in which $j - 1$ countries cooperate. A stable coalition is formed by j countries. Countries i and h are the marginal countries with respect to the stable coalition. Both i and h have an incentive to join the coalition (by definition of stability). Country i's most preferred outcome is the one in which it lets h cooperate. However, if country h does not act cooperatively, country i will choose to do so, in order to belong to the stable coalition (by definition of stability). Formally, this is implied by the following inequalities: $Q(j) > P(j+1) > P(j) > Q(j-1)$. The first and the last inequalities are implied by the stability of the j-coalition; $P(j+1) - P(j)$ holds by assumption (see subsections 2.2 and 2.3). Hence, non-cooperation is not the dominant strategy. This game is known as a *chicken game* (a game belonging to the class of coordination games). There are two equilibria (N, C) and (C, N), but all countries have an incentive to let the others cooperate.[9] The game has no dominant strategy; countries' attempts to choose non-cooperation, in order to let the others cooperate, may lead to the worst possible outcome (N, N). The cooperative outcome (C, C) is not Pareto optimal.

A stable coalition can be expanded by transfers to non-cooperating countries, provided some form of commitment takes place. The intuition behind this result is simple. Welfare transfers to non-cooperating countries decrease by T the payoff of the countries belonging to the j-coalition, preserving profitability (transfers are self-financed), but creating instability. The instability has to be dealt with by some form of commitment. As we anticipated in the introduction, the various forms of commitment we analyse are less demanding than the commitment by all players, assumed in cooperative games. In our framework, the

[9] The impasse can be solved by the introduction of asymmetries into the game. If countries have different preferences, technology or environmental endowment, it is possible to determine which countries are likely to form a coalition. For example, in the case of external commitment, countries with higher abatement costs are likely to finance emission reductions in countries with lower abatement costs, who therefore form the coalition. In the case of stable coalition commitment, countries in which environmental policy is part of a package of coordinated policies, or large countries that heavily affect the global environment, are more likely to commit themselves to cooperation, thus attracting other cooperators. If the game were repeated, countries with higher discount rate would be more likely to form a coalition.

commitment of only a fraction of the n countries can ensure the stability of wider coalitions, and can even lead to full cooperation.

Notice, moreover, that welfare transfers are indeed impossible and/or inefficient if they are based only on emission reductions. The simplest instrument to transfer wealth is probably cash. There are, however, other appropriate instruments, such as trade or debt policy and technology transfers.

3.2. Stable coalitions and best-reply functions

Over and above the previous comments, a number of specific questions deserve answers: Under what economic conditions does a stable coalition exist? How many countries belong to the stable coalition? Which form of commitment is likely to increase the number of countries belonging to the stable coalition? What is the size of the largest coalition in each different case?

To answer such questions, it is necessary to specify a particular form for the benefit and damage functions of the different countries. Under such restriction, the environmental literature does provide some helpful insights into stable coalitions: for example, Barrett (1991), Carraro and Siniscalco (1991) and Hoel (1992) show that stable coalitions far the protection of the environment exist under reasonable specifications of the benefit and damage functions. The same conclusion is reached in the cartel stability literature (see D'Aspremont et al., 1983; and Schmalensee, 1987, for example). More general results on the existence of stable coalitions in oligopolistic markets can be found in D'Aspremont and Gabszewicz (1986) and Donsimoni et al. (1986). These results prove that there are cases in which the metagame describing countries' interaction in environmental negotiations is not a Prisoner's dilemma. These works also show that stable coalitions are generally formed by a subset of all players of the game, and that this subset is often small. Only Barrett (1991) using a specific example, provides numerical simulations in which the number of cooperating countries approximates n.

In addition, there is one result which appears in most of the works on the international environment, but is seldom discussed: the pattern of interdependence among countries, as described by the slope of their best-reply function, is crucial for understanding the effectiveness of cooperative and non-cooperative emission control. The reason being that the more negative the slope of the best-reply functions, the larger the incentive to deviate from any coalition [i.e. $Q(j-1) - P(j)$ in the game]. Drawing from results proved in Carraro and Siniscalco (1991), we would like to clarify the role played by the slope of countries' best-reply function.

Let us first consider non-cooperative emission control. Countries that interact in a common environment with mutual externalities set their emissions by equating their own marginal benefit to marginal damage, given the emissions set by the other countries. In this context, country i's actual emissions are generally lower than emissions δ_i, which maximize its benefit function. Moreover, non-cooperative emissions are increasingly reduced as the slope of the best-reply functions become increasingly negative. The literature shows this is the case when the impact of foreign emissions for country *i* is high, the perceived damage is high, and the benefit (abatement cost) is low (on this point see Barrett, 1991; and Carraro and Siniscalco, 1991).

The reason that lies behind this result is intuitively simple. The best-reply functions reflect, inter alia, the marginal damage produced by foreign countries' emissions. If this marginal damage to country *i* is relevant, then the best non-cooperative response of country *i* to an expansion of foreign emission is an emission reduction. This reduction will be greater, the lower the benefit of domestic emissions, and vice versa. The difference between δ_I and the actual non-cooperative emissions of country *i* is therefore positively related to the slope of country *i*'s best-reply function.

Let us now consider cooperation. In this case, countries bargain over emission levels in order to achieve an optimal aggregate outcome, taking into account reciprocal externalities. As is well known, cooperation among all countries is profitable and optimal, but it is intrinsically undermined by free-riding. However, if the best-reply functions are orthogonal or near orthogonal, there is some scope for partial cooperation in the form of agreements among small groups of countries, which can be profitable and stable. The reason, again, has to do with free-riding behaviour, as reflected by the best reply function of a country which does not belong to the partial coalition. If this best-reply function is negatively sloped, the non-cooperating country will expand its emissions if the coalition restricts them, offsetting the effort of the cooperating countries. If, on the contrary, the best-reply functions are orthogonal or near orthogonal, the free-rider will simply enjoy the cleaner environment without paying for it, but will not offset the emission reduction by the cooperating countries.

Similarly, with negatively sloped best-reply functions, the number of countries is also crucial for the existence of stable coalitions. If *n* is large, a stable coalition is unlikely to exist, as a cooperative contraction by few countries is offset by the reactions of many others (see Barrett, 1991; and Carraro and Siniscalco, 1991).[10]

The above considerations suggest that, in environmental agreements, there is a sort of trade-off. When the best-reply functions are negatively sloped there is a high degree of interdependence. Likewise, non-cooperative emission control can lead to substantial emission reduction. But if one or more countries unilaterally or cooperatively reduces emissions, this contraction is offset by an expansion by the non-cooperating countries. This kind of interaction undermines all kinds of cooperation, as the free-riding behaviour implies a substantial loss for countries

[10] Notice that high perceived damage, low abatement cast, and high impact of foreign emissions are only necessary conditions for the negative slope of the best-reply functions. Separable damage functions imply orthogonal best-reply functions, whatever the other parameters.

who wish to cooperate. With an orthogonal or near-orthogonal best-reply function the situation is somehow reversed. Non-cooperative emission control leads to small emission reductions, but the scope for cooperation is now greater: if a number of countries cooperatively reduce their emissions, this reduction is not offset by free-riders, who simply enjoy a better environment but do not directly damage countries which cooperate. In this case, then, stable coalitions exist.

3.3. Transfers and commitments

The result that stable coalitions are formed by a subset of the n countries, and that this subset is often small, led us to consider strategies for expanding stable coalitions. Coalition expansions can be achieved through welfare transfers and some form of commitment. How much can a stable coalition be expanded? Which type of commitment leads to full cooperation?

Proposition 1 in section 2 holds for any form of the benefit and damage function: it shows that any attempt to expand a stable coalition by means of transfers is flawed without some form of commitment. This point, which may be relevant to the environmental policy debate, has already been discussed in section 2 and subsection 3.1.

We have attempted, therefore, to explore various forms of commitment in order to sustain expanded coalitions.

The ‘stable coalition commitment’ (Proposition 2) is the first to be discussed. Let us suppose a stable coalition exists. If all members of the coalition commit to cooperation, they can use the gains resulting from moving to a larger coalition to bribe other countries. How many countries can be induced to enter the coalition using this strategy?

The existing environmental literature is not very helpful in addressing these issues, as they are relatively new. One specific example we provide elsewhere (Carraro and Siniscalco, 1991), assuming symmetric countries and linear-quadratic benefit and damage functions, shows that the answer depends once more on the slope of the best-reply functions. With orthogonal (or near-orthogonal) best-reply functions, a stable coalition of three countries can induce four other countries to cooperate, irrespective of n. When the best-reply functions become negatively sloped, the possibility of bribing other countries gradually decreases to zero.

Besides the example mentioned above, we believe that this type of commitment can be relevant to capture some features of the current negotiation on global warming: why should a group of countries which already cooperate (say EC countries) commit and transfer ‘new and additional resources’ to other countries (say the LDCs)?

Sequential commitment, discussed in Proposition 3, is possibly less realistic, but shows how it is possible to reach wider coalitions. Countries belonging to a stable coalition commit to cooperation and bribe other countries. As any of the latter countries joins the coalition, it must commit in order to further expand the coalition. In this way, it is easy to show that, with orthogonal (or near-orthogonal) best-reply functions, sequential commitment leads to full cooperation, as the conditions on the stability of the coalition itself are removed by this form of commitment.

Full cooperation can also be achieved if about 60 percent of the *n* countries commit to cooperation (we are referring here to Proposition 4), while the external commitment of Proposition 5, again with orthogonal or near-orthogonal best-reply functions, can induce about 70 percent of the *n* countries to cooperate. The latter case can be helpful in understanding the recent proposal by some industrial countries (e.g. the Scandinavian countries) to subsidize environmental programmes in other countries (e.g. Eastern Europe; see Kaitala et al., 1992).

As the above discussion has already pointed out, if the best-reply functions are negatively sloped, all the results are generally less favourable. In particular, sequential commitment does not lead to full cooperation, and the impact of each type of commitment decreases as the slope of the best-reply functions increases (in absolute value).

As examples can provide only anecdotes and special results, we believe that a better understanding of environmental negotiations can only come from serious applied work. Only empirical work can justify alternative specifications of countries' interdependence. Only empirical work, moreover, can support an intuition we submit: while in the traditional case of common property goods (fisheries, pastures, forests, etc.) the payoff functions give rise to non-orthogonal best-reply functions, in the case of some global pollutants, e.g. CO_2 or CFC, the best-reply functions are probably orthogonal (or near orthogonal).[11]

4. Conclusions and Scope for Further Work

In the next few years the international protection of the environment will increasingly rely on international agreements, although they often involve substantial difficulties. To what extent can the proposed analysis be useful, and how can it be extended?

The analytical framework we proposed is highly simplified and the results obtained must be interpreted with great caution. However, given the difficulties and failures of many attempts to reach global agreements, it shows a promising route for research and policy analysis.

Firstly, the framework and the results show that it is rather sterile to study optimal agreements among all countries, if such agreements are profitable but

[11] Indeed, we can hardly think of any countries that expand their own CO_2 or CFC emissions in response to other countries' reductions. As the above discussion has suggested, this leaves room for cooperative agreements.

intrinsically unstable. The structure of such agreements can be useful as a benchmark, but it is very unlikely to be realized in practice. Secondly, they show that there is a full range of possible agreements among sovereign countries to protect the international environment. There are cases in which an effective protection can be obtained non-cooperatively. In other cases, an effective environmental protection can be reached through partial co-operation and transfers.

Coming to the scope for further work, a number of relevant issues have still to be addressed. First, it would be useful to provide a sort of taxonomy relating the various pollutants to appropriate damage functions. Only then would it be possible to contextualize policy analysis, obtaining meaningful results for each case. Secondly, it would be useful to re-appraise the instruments to implement cooperation. Emissions in many cases, are very difficult to monitor. The various economic instruments needed to implement an agreement, therefore, must be designed in a way that prevents cheating. So far, the literature has compared the various agreements in terms of efficiency, i.e. maximum profitability. Our analysis proposes another criterion: an instrument must be efficient, but it must also be effective in preventing or discouraging free-riding. In other words, it must also be designed to promote the stability of the agreements.

Finally, three extensions should be attempted:

(i) Asymmetric information should be introduced. Countries' preferences can hardly be observed. If we remove the assumption of complete information, each country that is induced to enter a coalition would be tempted to overstate the damage and claim for greater incentives. The solution to this problem is to embody an appropriate information or self-selection premium in the transfer to each country that enters the coalition.

(ii) It would be relevant to introduce some asymmetries into the game (by assuming for example different damage and abatement costs across countries), in order to evaluate coalition profitability and stability when countries have different incentives to join it.

(iii) The benefit function should account for the interaction between environmental variables and the policy instruments designed to carry on transfers. This prevents the analysis of environmental policy as such, but implies its integration into a wider analysis which accounts for other economic variables in each country's payoff.

A last point concerns institutions. In the present setting there is not an institution that has the authority to impose supernational regulations on countries and regions. Stable and expanded coalitions can be seen as a first step towards such institutions, as the member countries would find it costly to individually perform the transfer, the monitoring, and the enforcement activities which are usually associated with the management of a cooperative agreement.

References[12]

Barrett, S. (1991), 'The paradox of international environmental agreements', mimeo, London Business School.

Barrett, S. (1992), 'International environmental agreements as games', in R. Pethig (ed.), *Conflicts and Cooperation in Managing Environmental Resources*, Berlin: Springer Verlag, pp. 18–33.

Binmore, K., A. Rubinstein and A. Wolinski (1986), 'The Nash bargaining solution in economic modelling', *Rand Journal of Economics* **17**, 176–188.

Bohm, P. (1990), 'Efficiency aspects of imperfect treaties on global public bads: Lessons from the Montreal protocol', mimeo.

Carraro, C. and D. Siniscalco (1991), 'Transfers and commitments in international negotiations', paper prepared for the ESF task force 3 on environmental economics, in K.G. Maler (ed.) (1993), *International Environmental Problems: An Economic Perspective,* Dordrecht: Kluwer Academic Publishers.

D'Aspremont, C.A. and J.J. Gabszewicz (1986), 'On the Stability of collusion', in G.F. Matthewson and J.E. Stiglitz (eds.), *New Developments in the Analysis of Market Structure,* New York: Macmillan, pp. 243–264.

D'Aspremont, C.A., A. Jacquemin, J.J. Gabszewicz and J. Weymark (1983), 'On the stability of collusive price leadership', *Canadian Journal of Economics* **16**, 17–25.

Donsimoni, M.P., N.S. Economides and H.M. Polemarchakis (1986), 'Stable cartels', *International Economic Review* **27**, 317–327.

Hardin, G. (1968), 'The tragedy of commons', *Science* **162**, 1243–1248.

Hardin, G. and J. Baden (1977), *Managing the Commons,* New York: Freeman and Co.

Hoel, M. (1992), 'International environment conventions: The case of uniform reductions of emissions', *Environmental and Resource Economics* **2**, 141–160.

Kaitala, V., M. Pojola and O. Tahvonen (1992), 'Trans-boundary air poliution and soil acidification: A dynamic analysis of acid rain game between Finland and the USSR', *Environmental Resource Economics* **2**, 161–182.

Maler, K.G. (1989), 'The acid rain game', in H. Folmer and E. Ireland (eds.), *Valuation Methods and Policy Making in Environmental Economics,* New York: Elsevier, pp. 188–205.

Newbery, D.M. (1991), 'Acid rains', *Economic Policy* **1**, 42–88.

Nitze, W.A. (1990), *The Greenhouse Effect: Formulating a Convention,* London: The Royal Institute of International Affairs.

Ostrom, E. (1990), *Governing the Commons,* Cambridge: Cambridge University Press.

Schmalensee, R. (1987), 'Competitive advantage and collusive optima', *International Journal of Industrial Organisation* **5**, 351–367.

[12] [Editor's note] The reference Maler (1993) in Carraro and Siniscalco (1991), which was originally cited as forthcoming, has been updated.

Sequential Formation of Coalitions in Games with Externalities and Fixed Payoff Division

Francis Bloch

Paper presented at the CTN Workshop 1995. Reprinted from *Games and Economic Behavior* **14**(1), May 1996, 90–123, with permission from Elsevier.

This paper analyzes a sequential game of coalition formation when the division of the coalitional surplus is fixed and the payoffs are defined relative to the whole coalition structure. Gains from cooperation are represented by a valuation which maps coalition structures into payoff vectors. I show that any core stable coalition structure can be attained as a stationary perfect equilibrium of the game. If stationary perfect equilibria may fail to exist in general games, a simple condition is provided under which they exist in symmetric games. Furthermore, symmetric stationary perfect equilibria of symmetric games generate a coalition structure which is generically unique up to a permutation of the players. A general method for the characterization of equilibria in symmetric games is proposed and applied to the formation of cartels in oligopolies and coalitions in symmetric majority games.

1. Introduction

Since the publication of *Theory of Games and Economic Behavior*, the study of endogenous formation of coalitions has been one of the most intriguing and

This paper is based on the second chapter of my Ph.D. dissertation at the University of Pennsylvania. I am grateful to my advisor, Beth Allen, for her help and encouragement. Discussions with Elaine Bennett, Ken Binmore, Roger Lagunoff, George Mailath, Andrew Postlewaite, Ariel Rubinstein and Sang-Seung Yi were extremely valuable. I am particularly indebted to Debraj Ray and Rajiv Vohra for pointing out an error in an earlier version of this paper. I am also grateful to participants at seminars at Boston University, Brown, Columbia, Stony Brook and the European Meeting of the Econometric Society in Brussels for their comments. Detailed comments by an associate editor greatly improved the quality of the paper. This work was partly supported by a Sloan Foundation Doctoral Dissertation Fellowship.

challenging problems open to game theorists. Many solution concepts such as Von Neumann and Morgenstern's stable sets (Von Neumann and Morgenstern, 1944) and Aumann and Maschler's bargaining set (Aumann and Maschler, 1964) were in fact primarily designed as ways to solve the problem of joint determination of a coalition structure and the allocation of the coalitional surplus among coalition members. While these approaches proved fruitful in the study of many situations of cooperation, they mostly rely on the assumption that gains from cooperation can be defined independently of the coalitions formed by external players.[1] Using the terminology introduced by Shubik (1982), cooperative game theory has focused on games with orthogonal coalitions which are well-suited to situations of pure competition but fail to capture the effects of externalities among coalitions. The objective of this paper is to propose a model of formation of coalitions in nonorthogonal games where payoffs depend on the whole coalition structure.

The presence of externalities among coalitions introduces a new difficulty in the study of endogenous coalition formation. When players decide to form a coalition, they must take into account the reaction of external players to the formation of the coalition. The sequential model analyzed in this paper addresses this problem by explictly describing a procedure in which individual players, when deciding to form a coalition, consider the consequences of their actions on the behavior of the other players. However, to keep the analysis tractable and concentrate on the role played by externalities on the formation of the coalition structure, I do not model the allocation of the coalitional surplus among members of a coalition, and assume instead that the coalitional worth is distributed according to a fixed sharing rule. Gains from cooperation are then represented by a valuation which maps coalition structures into vectors of individual payoffs.

Arguably, the assumption that payoffs are determined by a fixed rule is very restrictive and may seem a high price to pay for allowing externalities among coalitions. But valuations arise naturally in two distinct categories of economic models and the study of coalition formation in games represented by a valuation may appear fruitful in the resolution of these models.

First, valuations are considered in the models of coalition formation studied by Myerson (1978), Shenoy (1979), Hart and Kurz (1983) and Aumann and Myerson (1988). In these models, the formation of coalitions is viewed as a two-stage process where players form coalition in the first stage and decide on the allocation of the coalitional surplus, given a fixed coalition structure, in the second stage. Hence, at the time coalitions are formed, players evaluate the payoffs they receive in each coalition structure according to a fixed rule.

The exact characterization of the rule employed in the second stage depends

[1] Two important exceptions are Thrall and Lucas (1963)'s study of games in partition function form and Aumann and Drèze (1974)'s analysis of games with fixed coalition structures.

on the situations considered in the different models. In Myerson (1978)'s threats and settlement game, the fair settlement function assigns to each collection of coalitions (not necessarily a coalition structure) a unique vector of payoffs.

Shenoy (1979) uses as an evaluation rule Aumann and Drèze (1974)'s extension of the Shapley Value to games with fixed coalition structures. In Hart and Kurz (1983)'s analysis, players evaluate coalition structures according to a different extension of the Shapley Value first analyzed by Owen (1977). In Aumann and Myerson (1988)'s study of formation of links among players, the valuation used is Myerson (1977)'s extension of the Shapley Value to games with cooperation graphs of players.[2]

Second, valuations emerge in various applications of Game Theory to Industrial Organization and Public Economics involving competing coalitions of economic agents. The study of the formation of cartels in oligopolies leads to a natural definition of a valuation representing, for each cartel structure, the payoffs obtained by the firms belonging to the different cartels.[3] Similarly, the formation of associations of firms which agree to share some common resource but behave as competitors on the market can be analyzed with the use of a valuation.[4] The analysis of the provision of local public goods in a spatial setting where members of a community can benefit from the public goods provided in neighboring communities also requires the use of a valuation.[5] As a final example, the formation of customs unions allowing national firms to compete in a market characterized by the existence of different customs unions also leads to the definition of a valuation.

Cooperative solution concepts for games represented by a valuation were introduced by Shenoy (1979) and Hart and Kurz (1983) in their models of endogenous coalition formation.[6] To predict which coalitions will be formed, they propose different definitions of stability of coalition structures.[7] The variety of stability concepts accounts for the fact that, in games described by a valuation, the payoffs obtained by members of a blocking coalition depend on the reaction of the external players. The solution concepts range from the core stability concept, which supposes a very optimistic conjecture about the reaction of the external players since players deviate if there exists a coalition structure in which they are better off to the

[2] In Myerson (1978) and Hart and Kurz (1983), the emphasis is put on the axiomatic derivation of a reasonable valuation rather than on the first stage game of coalition formation. This paper, by contrast, focuses on the game of coalition formation.

[3] Salant et al. (1983) were the first to point out in a simple model the problems of cartel formation in oligopolies. Yi and Shin (1995) contains a very complete description of the derivation of the valuation in the cartel problem.

[4] The study of associations of firms, which can be interpreted as Research Joint Ventures or standardization committees, is taken up in a distinct paper (Bloch, 1995).

[5] Guesnerie and Oddou (1981) analyze the provision of local public goods in a model with orthogonal coalitions but discuss the role of externalities among communities.

[6] Hart and Kurz (1983) analyze strong equilibria of a noncooperative game where players simultaneously announce coalitions.

[7] Other concepts of stability of coalition structures are surveyed in Greenberg (1995).

α stability concept which is based on pessimistic conjectures since a coalition only deviates when it is guaranteed to obtain a higher payoff independently of the reaction of the other players. The study of stable coalition structures raises three important difficulties. First, the definitions of stability rely on ad hoc assumptions on the behavior of the other players after a coalition has deviated. Second, all definitions of stability assume that external players react to the formation of a coalition in a myopic way. Hence, when a coalition forms, its members do not take into account the final result of their decisions but only the immediate reaction of the other players. Finally, even the less restrictive definition of stability (α stability) may not be useful, since α stable coalition structures fail to exist in situations which are not easily characterized. (Hart and Kurz (1984) give an example of a game without stable structure which is otherwise well-behaved.)

By contrast, in this paper, I explicitly model the formation of coalitions as a noncooperative sequential process in the spirit of Rubinstein (1982)'s alternating offers bargaining game and its extensions to n players by Selten (1981) and Chatterjee et al. (1993). Players are ranked according to an exogenous rule of order. The first player starts the game by proposing the formation of a coalition. If all prospective members accept the proposal, the coalition is formed. If one player rejects the proposal, she becomes the initiator in the next round. The important feature of the game is that, once a coalition is formed, the game is only played among the remaining players and that established coalitions may not seek to attract new members nor break apart. Hence, by agreeing to group in a coalition, players commit to stay in that coalition.

I restrict my attention to stationary strategies and establish the following properties of stationary perfect equilibria. I first show that, if the game always admits a subgame perfect equilibrium, stationary perfect equilibria may fail to exist. A sufficient condition for the game to admit a stationary perfect equilibrium is that the valuation and all its restrictions to smaller sets of players admit core stable structures. Furthermore, any core stable coalition structure can be reached as a stationary perfect equilibrium of the extensive form game of coalition formation, provided that the set of stationary perfect equilibria is nonempty. I then study the restricted class of symmetric games where all players are ex ante identical. In this class of games, using a result due to Ray and Vohra (1995), I provide a simple condition under which symmetric stationary perfect equilibria exist, and I show that the coalition structures they generate are generically unique up to a permutation of the players. Furthermore, I provide a general method for the characterization of the coalition structures generated by symmetric stationary perfect equilibria in symmetric games. This method is used to derive equilibrium coalition structures in two situations: the formation of cartels in a symmetric oligopoly and the symmetric majority games discussed by Hart and Kurz (1984).

The game analyzed here is similar to games of coalition formation proposed by Selten (1981), Chatterjee et al. (1993), Moldovanu (1992) and Winter (1993) in the context of games in coalitional form. The games they analyze have the same sequence of moves as the one described above. The crucial difference between their games and mine stems from differences in the action spaces. By fixing the division of the payoffs, I restrict the actions of the agents to announcements of coalitions whereas they study a more general framework where players announce both a coalition and the division of the coalitional worth. A further difference is due to the underlying specification of gains from cooperation since they do not allow for externalities among coalitions. Given these differences, the results they obtain are not directly comparable to mine.

Different extensive form procedures of coalition formation in games represented by a valuation were proposed by Aumann and Myerson (1988) and Shin and Yi (1995). The procedure in Aumann and Myerson (1988) is defined for games where players evaluate cooperation graphs rather than coalition structures. The particular feature of cooperation graphs where coalition members need not unanimously agree to admit new members leads them to define a game where links can be formed at any stage. This approach cannot easily be applied to situations where gains from cooperation accrue when coalitions are formed, rather than bilateral links among players. Yi and Shin (1995) analyze games based on a 'matching procedure'. Players announce coalitions and coalitions are formed whenever all its members have made identical announcements. In general, the equilibria they obtain are very different from the equilibria of the infinite horizon game analyzed in this paper.

The paper is organized as follows. The game of sequential formation of coalitions is introduced and the equilibrium concept defined in Section 2. In Section 3, I analyze the relations between stationary perfect equilibria and stability concepts for coalition structures in games described by a valuation. Section 4 is devoted to the analysis of symmetric games. I present applications of the model to the formation of cartels in oligopolies and of coalitions in symmetric majority games in Section 5. My concluding remarks and some directions for future research appear in Section 6.

2. Sequential Formation of Coalitions

In this section, I introduce the sequential game of coalition formation and the equilibrium concept that I will use. The set of players is denoted N, with cardinality n. The index i will refer to the players. A *coalition* T is a nonempty subset of players. A *coalition structure* π is a partition on the set N. The set of all coalition structures is denoted by Π. For any subset K of N, the set of partitions on K is denoted Π_K with typical element π_K.

Gains from cooperation are described by a *valuation* v which maps the set of coalition structures Π into vectors of payoffs in $\Re^n$. The component $v_i(\pi)$ denotes the payoff obtained by player i if the coalition structure π is formed. I assume that payoffs are normalized so that any player, by opting to leave the game can get a strictly positive payoff. Formally, $\forall i \in N$, $\min_{\pi \supset \{\{i\}\}} v_i(\pi) > 0$.

A *rule of order* ρ is an ordering of the players, which is used to determine the order of moves in the sequential game of coalition formation.

The sequential game of coalition formation is defined by the exogenous specification of the valuation v and of the rule of order ρ. To emphasize this dependence, I denote the game of coalition formation by $\Gamma(v, \rho)$.

The game $\Gamma(v, \rho)$ proceeds as follows. The first player according to the rule of order ρ starts the game by proposing the formation of a coalition T to which she belongs. Each prospective member responds to the proposal in the order determined by ρ. If one of the player rejects the proposal, she must make a counteroffer and propose a coalition T' to which she belongs. If all members accept, the coalition is formed. All members of T then withdraw from the game, and the first player in $N \setminus T$ starts making a proposal.[8]

This game describes in the simplest way a procedure where coalitions are formed *in sequence*. The main characteristic of the game is that, once a coalition has been formed, the game is only played among the remaining players. The extensive form thus embodies a high degree of commitment of the players. When players agree to join a coalition, they are bound to remain in that coalition. They can neither leave the coalition nor propose to change the coalition at later stages of the game. Figure 1 depicts the extensive form of the game with three players.

A *history* h^t at date t is a list of offers, acceptances and rejections up to period t. At any point in the game $\Gamma(\rho, v)$, a history h^t determines

- a set $\hat{K}(h^t)$ of players who have already formed coalitions
- a coalition structure $\pi_{\hat{K}(h^t)}$ formed by the players in $\hat{K}(h^t)$
- an ongoing proposal (if any) $\hat{T}(h^t)$
- a set of players who have already accepted the proposal
- a player who moves at period t .

Player i is called *active* at history h^t is it is her turn to move after the history h^t. The set of histories at which player i is active is denoted H_i. A *strategy* σ_i for player i is a mapping from H_i to her set of actions, namely

$$\sigma_i(h^t) \in \{\text{Yes, No}\} \quad \text{if } \hat{T}(h^t) \neq \varnothing$$

$$\sigma_i(h^t) \in \{T \subset N \setminus \hat{K}(h^t), i \in T\} \qquad \text{if } \hat{T}(h^t) = \varnothing$$

When $\hat{T}(h^t) \neq \varnothing$, player i is a respondent to the offer $\hat{T}(h^t) = \varnothing$ and she can

[8] Each time a coalition T is proposed, the order of responses is fixed by ρ independently of the history or the identity of the proposer. Hence, for example, if player 2 proposes the formation of a coalition $\{1, 2, 3\}$, player 1 responds first and player 3 responds after player 1.

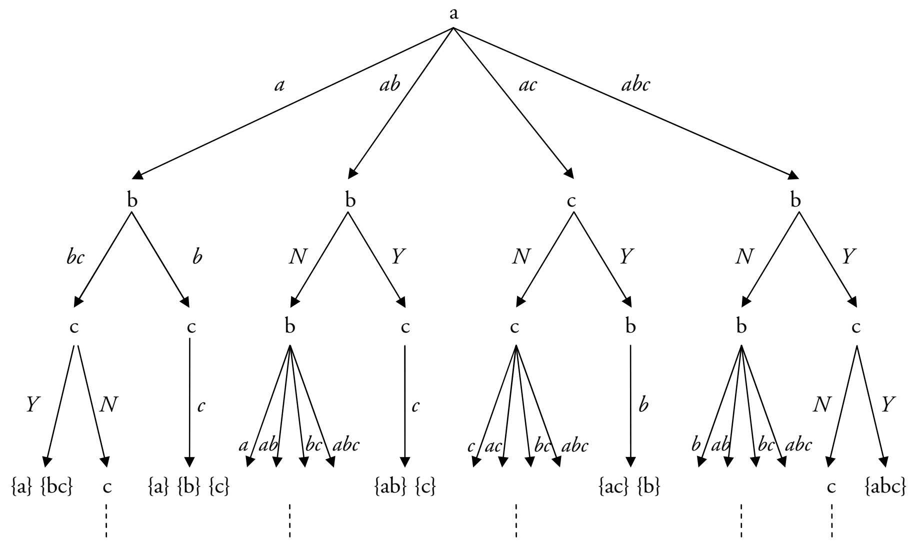


Figure 1. The game Γ

choose to accept or reject it. If $\hat{T}(h^t) \neq \emptyset$, either a coalition has just formed and player i is the first player in $N \setminus \hat{K}(h^t)$ according to the rule of order ρ, or player i has just rejected an offer. In both cases, it is her turn to propose a new coalition T which must be a subset of the remaining players to which she belongs.

I restrict my attention to strategies which only depend on the payoff-relevant part of the history. For a player i active at history h^t, the only payoff-relevant features of the history are the set K of players who left the game, the partition π_K representing the coalitions they formed and the current offer T. In particular, the set of players who have already accepted the offer T is uniquely determined by the rule of order ρ.

A strategy σ_i is *stationary* if it only depends on the state $s = (K, \pi_K, T)$ where K is a (possibly empty) subset of N, π_K is a partition of K and T is a (possibly empty) subset of $N \setminus K$. Formally, letting $\mathcal{T}(i, K)$ define the collection of subsets of $N \setminus K$ to which player i belongs, a stationary strategy is a mapping from the set of states at which player i is active, S_i, to a set of actions, where

$$\sigma_i(K, \pi_K, T) \in \{\text{Yes, No}\} \text{ if } T \neq \emptyset$$

$$\sigma_i(K, \pi_K, \emptyset) \in \mathcal{T}(i, K)$$

Any strategy profile $\sigma = \{\sigma_i\}_{i \in \mathrm{N}}$ determines an outcome $(\pi(\sigma), t(\sigma))$ of the game. If the game ends in a finite number of periods, $\pi(\sigma)$ is a coalition structure on the set N, and $t(\sigma)$ is the period at which the agreement has been reached.

I assume that players do not discount the future. In the case of an infinite play of the game, players who have not formed a coalition receive a payoff of zero.

More precisely, suppose that a subset $N \setminus K$ of the players does not reach an agreement in a finite number of periods. Payoffs are then given by

$$v_i(\pi(\sigma)) = 0 \qquad \text{for all players in} \quad N \setminus K$$

$$v_i\ (\pi(\sigma)) = \max_{\pi_K \subset \pi} v_i(\pi) \qquad \text{for all players in} \quad K.$$

Definition 2.1. A ***subgame perfect equilibrium*** σ^* is a strategy profile such that $\forall i \in N$, $\forall h^t \in H_i$, $\forall \sigma_i$, $v_i(\pi(\sigma_i^* (h^t), \sigma_{-i)}^*)) \geq v_i(\pi(\sigma_i\ (h^t), \sigma_{-i)}^*)$.

Definition 2.2. A ***stationary perfect equilibrium*** σ^* is a subgame perfect equilibrium where $\forall i \in N$, σ_i^* is a stationary strategy.

A coalition structure π generated by a subgame perfect equilibrium is called an *equilibrium coalition structure* (*ECS*). Coalition structures generated by stationary perfect equilibria are called *stationary equilibrium coalition structures* (*SECS*). The set of stationary equilibrium coalition structures is denoted $SECS(v, \rho)$.

Remark 2.3. Since every player obtains a higher payoff by leaving the game than by disagreeing forever, an infinite play of the game cannot be part of a subgame perfect equilibrium. Hence, the concept of an equilibrium coalition structure is well defined.

The payoffs of the game described above are not continuous at infinity. Hence the existence of a subgame perfect equilibrium is not guaranteed. To circumvent this difficulty, I first show that any subgame perfect equilibrium of the game with sufficiently high discounting is a subgame perfect equilibrium of the game $\Gamma(v, \rho)$. To be more precise, let $\Gamma_\delta(v, \rho)$ denote the game where strategies and moves are defined as above but payoffs are given by: $v_i(\sigma) = \delta_i^{t(\sigma)} v_i(\pi(\sigma))$.

Proposition 2.4. *There exists* $\underline{\delta} \in (0, 1)$ *such that, if* $\forall i$, $\delta_i > \underline{\delta}$, *any sub-game perfect equilibrium of* $\Gamma_\delta(v, \rho)$ *is a subgame perfect equilibrium of* $\Gamma(v, \rho)$.

Proof. Observe first that, since Π is finite, the set of payoffs of the game, $v(\Pi)$ is finite. Hence, the set of possible coalition structures formed in $\Gamma_\delta(v, \rho)$ is finite. In particular, this implies that, as δ varies continuously from 0 to 1, the strategy profiles of the game can only lead to a finite number of coalition structures. Hence, there exists a $\underline{\delta}$ such that, for all δ, $\delta' > \underline{\delta}$, if σ^* is a subgame perfect equilibrium of $\Gamma_\delta(v, \rho)$, then σ^* is a subgame perfect equilibrium of $\Gamma_{\delta'}(v, \rho)$.

Consider now $\delta' > \underline{\delta}$, and let σ^* be a subgame perfect equilibrium of $\Gamma_{\delta'}(v, \rho)$. Then, for any player i, any history h^t in H_i, any strategy σ_i and any $\delta \in [\delta', 1)$,

$$\delta_i^{t(\sigma_i^*(h^t),\sigma_{-i}^*)}\, v_i(\pi(\sigma_i^*; \sigma_{-i}^*)) \geq \delta_i^{t\,(\sigma_i(h^t),\sigma_{-i}^*)}\, v_i(\pi(\sigma_i, \sigma_{-i}^*)).$$

Taking limits as δ goes to 1,

$$v_i(\pi(\sigma_i^*(h^t), \sigma_{-i}^*)) \geq v_i(\pi(\sigma_i\,(h^t), \sigma_{-i}^*)).$$

Hence, σ^* is a subgame perfect equilibrium of $\Gamma(v, \rho)$. ■

Corollary 2.5. *For any valuation v and any rule of order ρ, there exists a subgame perfect equilibrium of the game $\Gamma(v, \rho)$.*

Proof. Fix a $\delta > \underline{\delta}$. The game $\Gamma_\delta(v, \rho)$ is a finite action game of perfect information and is continuous at infinity. Hence, by a result of Fudenberg and Levine (1983) (Corollary 4.2, p. 262), the game $\Gamma_\delta(v, \rho)$ has a subgame perfect equilibrium. From Proposition 2.4, any subgame perfect equilibrium of $\Gamma_\delta(v, \rho)$ is a subgame perfect equilibrium of $\Gamma(v, \rho)$. ■

By imposing stationarity, I require that strategies only depend on the payoff-relevant part of the history. In the framework analyzed here, the payoff-relevant part of the history is summarized by the state *s* characterizing the coalition structure formed by the previous players and the ongoing offer. Chatterjee et al. (1993) and Moldovanu (1992) show that, when players bargain over the division of the coalitional worth, the set of nonstationary perfect equilibria may be very large, and stationarity is a useful restriction to refine the set of subgame perfect equilibria. A striking aspect of the game analyzed here is that *stationary perfect equilibria may fail to exist*. This point is illustrated by the following example.

EXAMPLE 2.6. $N = \{a, b, c\}$, and ρ defines $a < b < c$.

π	$v_a(\pi)$	$v_b(\pi)$	$v_c(\pi)$
$a\|b\|c$	1	1	1
$ab\|c$	3	2	1
$ac\|b$	2	1	3
$a\|bc$	1	3	2
abc	1	1	1

In this example, player *a* wants to form a coalition with player *b,* player *b* with player *c,* and player *c* with player *a.*

To show that the game $\Gamma(v, \rho)$ does not admit any stationary equilibrium coalition structure, observe first that the three coalition structures $\{\{a, b, c\}\}$, {{*a*}, {*b*}, {*c*}} and {{*a*}, {*b, c*}} cannot be supported by any equilibrium since player *a* would benefit from deviating and offering the formation of the coalition {*a, c*} which player *c* would accept. The two other coalition structures {{*a, b*}, {*c*}} and {{*a, c*}, {*b*}} can be supported by equilibria in nonstationary strategies but not by a stationary perfect equilibrium. For {{*a, b*}, {*c*}} to be supported by a stationary perfect equilibrium, it must be that player *c* rejects the offer {*b, c*}. But, in equilibrium, player *c* will only reject the offer {*b, c*} if player *a* accepts the offer {*a, c*}. By stationarity, player *b* accepts the offer {*a, b*} irrespective of the history of rejections which have preceded it. Hence, since player *b* always accepts the offer {*a, b*}, player *a* cannot accept the offer {*a, c*}. Similarly, the coalition structure {{*a, c*}, {*b*}} is only supported by a strategy prescribing that player *b* rejects the offer {*a, b*}, implying that player *c* accepts the offer {*b, c*}. Since, by stationarity, player *a* always accepts the offer {*a, c*}, player *c* should reject the offer {*b, c*}. Hence, the game $\Gamma(v, \rho)$ does not admit any stationary perfect equilibrium.

However, the coalition structures {{*a, b*}, {*c*}} and {{*a, c*}, {*b*}} can be supported by equilibria in nonstationary strategies.[9] To support these coalition structures as equilibria, one only needs to allow players to condition their actions on the number of times they have received an offer. Consider first the coalition structure {{*a, b*}, {*c*}} and the following strategies. Player a always accepts the offer {*a, b*} and proposes {*a, b*}. She rejects {*a, b, c*} and accepts {*a, c*} when, in the history h^t, she has made the offer {*a, b*} to player *b* an *odd* number of times. Player *b* accepts {*b, c*} and proposes {*b, c*}. She rejects {*a, b, c*} and only accepts {*a, b*} if, in the history h^t, the offer {*a, b*} has been made by player *a* an *odd* number of times. Player *c* accepts {*a, c*} and proposes {*a, c*}. She rejects {*a, b, c*} and only accepts {*b, c*} if, in the history h^t, player *a* has made the offer {*a, b*} an *even* number of times. These strategies form a subgame perfect equilibrium of the game (in nonstationary strategies), and are depicted in Figure 2. A strategy profile supporting the coalition structure {{*a, c*}, {*b*}} can be constructed in a similar way.

In Example 2.6, the three players play a symmetric role. Hence, no change in the rule of order can guarantee the existence of a stationary perfect equilibrium. Moreover, Example 2.6 is generic, since the nonexistence of a stationary perfect equilibrium is robust to small variations of the valuation. Nonexistence of stationary perfect equilibria is thus a robust phenomenon in games with more than three players.

Note however that the nonexistence of a stationary perfect equilibrium in pure strategies in Example 2.6 is linked to the fixed sharing rule. If players were allowed to bargain freely over the worth of the coalition in a game with transferable utility, the nonexistence result would disappear.

[9] These strategies are closely related to strategies constructed by Shaked to support any division of the payoffs in a three-person bargaining game (Sutton, 1986).

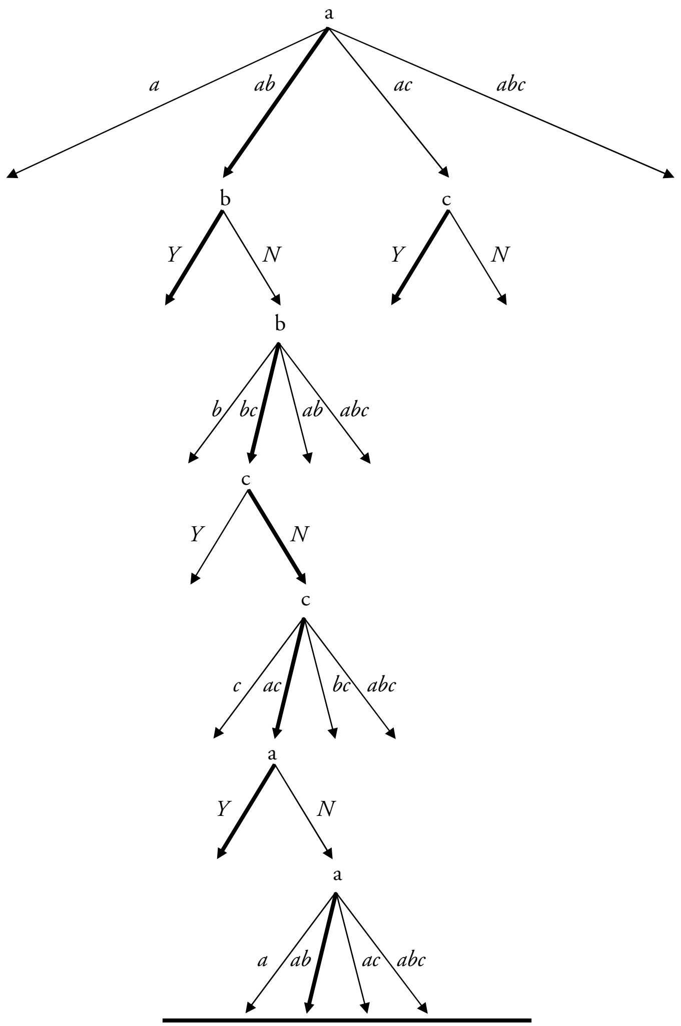


Figure 2. Nonstationary equilibrium strategies supporting the coalition structure $\{\{a, b\}, \{c\}\}$

The central feature of Example 2.6 is the disagreement among players over the coalitions which should be formed. A similar problem was noted by Shenoy (1979) in Apex games, where a single big player faces a number of small players (Example 7.5, p. 150). The preferred coalition for the big player is the grand coalition, since it offers her the possibility of diluting the power of the small players. Small players, on the other hand, would rather form a two-member coalition with the big player. This disagreement among players about the coalition which should be formed leads, as in Example 2.6, to the nonexistence of a stationary perfect equilibrium. This suggests that a sufficient condition for the existence of an equilibrium coalition structure is high degree of unanimity among players about the coalitions they wish to form. While this point is not pursued here, the class of symmetric games analyzed in Section 4 provides an example of games where players unanimously agree on the coalitions they want to belong to.

3. Stable Coalition Structures

In this section, I compare the equilibrium coalition structures with coalition structures satisfying cooperative concepts of stability. Concepts of stability in games with externalities require a specification of the reaction of external players to the formation of a coalition, and different assumptions on the behavior of external players give rise to different definitions of stability. Kurz (1988) distinguishes five models of reaction of the external players. The core stability concept, first introduced by Shenoy (1979), is based on the following dominance relation. A coalition structure π dominates a coalition structure π' if there exists a coalition in π whose members receive strictly higher payoffs than in π'. A coalition structure is called *core stable* if it belongs to the core of the dominance relation. In effect, this definition of stability is very restrictive, since it assumes that, when a group of players deviate, they consider that external players react in such a way as to maximize the payoff of deviating players.

Hart and Kurz (1983) propose four models of reaction of the external players. In the γ model, coalitions which are left by some members dissolve. In the δ model, members of coalitions which lose members remain together and form smaller coalitions. The last two stability concepts are based on the β and the α cores.[10] In the β model, a group K of players deviates if, for any possible reaction of the external players, namely any coalition structure $\pi_{N \setminus K}$ of $N \setminus K$, there exists a coalition structure of K, π_K, such that all members of K are better off in the new coalition structure $\pi = \pi_{N \setminus K} \cup \pi_K$. In the α definition, a group K of players deviates if there exists a coalition structure π_K such that, whatever the reaction of the external players, members of K are better off forming the coalition structure π_K.

Letting, for any fixed valuation v, the sets of Core stable, γ stable, δ stable, β stable and α stable coalition structures be denoted by $CC(v)$, $C\gamma(v)$, $C\delta(v)$, $C\beta(v)$ and $C\alpha(v)$, the following lemma is easily established.[11]

Lemma 3.1. *For any valuation v, $CC(v) \subset (C\gamma(v) \cup C\delta(v)) \subset C\beta(v) \subset C\alpha(v)$.*

I will focus here on the two extreme concepts of core and α stability.[12]

[10] See Aumann (1967) for a complete description of the α and β core concepts.

[11] Hart and Kurz (1983) derive the last three inclusions of the Lemma. The first inclusion is immediate, once one reinterprets the core stability concept in terms of reaction of the external players to the deviation of a group of players.

[12] The absence of coincidence between α stable structures and equilibrium coalition structures can be extended to the intermediate concepts of β, γ, and δ stability.

Formally, a coalition structure π is *core stable* if there does not exist a coalition K and a coalition structure π' such that $K \in \pi'$ and $\forall i \in K,\ v_i(\pi') > v_i(\pi')$. A coalition structure π is *α stable* if there does not exist a coalition K and a partition π'_K on K such that, $\forall i \in K, \forall \pi_{N \setminus K} \in \Pi_{N \setminus K}, v_i(\pi'_K \cup \pi_{N \setminus K}) > v_i(\pi)$.

The next proposition shows that, when the set of stationary equilibrium coalition structures is nonempty, it contains the set of core stable structures.

Proposition 3.2. *Assume that there exists a rule of order ρ such that $SESC(v, \rho) \neq \emptyset$. Then $CC(v) \subset SECS(v, \rho)$.*

Proof. Let $\tilde{\rho}$ denote one rule of order for which $SECS(v, \tilde{\rho}) \neq \emptyset$. Let $\tilde{\pi}$ denote a coalition structure in $CC(v)$. To prove the proposition, I construct a stationary perfect equilibrium $\tilde{\rho}$ of the game $\Gamma(v, \tilde{\rho})$ such that $\pi(\tilde{\sigma}) = \tilde{\pi}$. I denote by $T(i)$ the coalition to which player i belongs in the coalition structure $\tilde{\pi}$. A partition π_K of a subset K of the players is called a *subpartition* of $\tilde{\pi}$ if it is formed by the union of elements of $\tilde{\pi}$. The set of all subpartitions of $\tilde{\pi}$ is denoted $Sub(\tilde{\pi})$. Pick a stationary perfect equilibrium $\tilde{\tilde{\sigma}}$ of the game $\Gamma(v, \tilde{\rho})$. A stationary strategy $\tilde{\sigma}_i$ for player i is then constructed as follows.

Assume that a subset K of players, where $i \notin K$, has already formed a coalition structure π_K.

$$\begin{aligned}
&\text{If } \pi_K \notin Sub(\tilde{\pi}), && \tilde{\sigma}_i(K, \pi_K, \cdot) = \tilde{\tilde{\sigma}}_i\,(K, \pi_K, \cdot) \\
&\text{If } \pi_K \in Sub(\tilde{\pi}), && \tilde{\sigma}_i(K, \pi_K, \phi) = T(i) \\
& && \tilde{\sigma}_i(K, \pi_K, T(i)) = \text{Yes} \\
& && \tilde{\sigma}_i(K, \pi_K, T') = \text{Yes} \quad \text{if} \quad v_i(\pi(T')) > v_i(\tilde{\pi}) \\
& && \tilde{\sigma}_i(K, \pi_K, T') = \text{No} \quad \text{if} \quad v_i(\pi(T')) \leq v_i(\tilde{\pi})
\end{aligned}$$

where $\pi(T')$ is the coalition structure generated by $\tilde{\tilde{\sigma}}$ after the coalition T' has been formed.

The strategy $\tilde{\sigma}$ prescribes that player i follows her part of a stationary perfect equilibrium $\tilde{\tilde{\sigma}}$ if a coalition structure π_K off the equilibrium path has been formed, and that she forms the coalition $T(i)$ otherwise.

It remains to check that $\tilde{\sigma}$ is a subgame perfect equilibrium of the game $\Gamma(v, \tilde{\rho})$. Observe first that, since $\tilde{\tilde{\sigma}}$ is a stationary perfect equilibrium profile, the strategy profile $\tilde{\sigma}$ is a subgame perfect equilibrium if a coalition structure off the equilibrium path has been formed. Suppose now that the previous players have formed a coalition structure π_K in $Sub(\tilde{\pi})$. To check that $\tilde{\sigma}$ is a subgame perfect equilibrium on the equilibrium path, consider the possible deviations for player i.

Player i can deviate by announcing a coalition structure $T' \neq T(i)$ when it is her turn to announce a coalition. However, since $\tilde{\pi}$ is a core stable structure, there exists a player j in T' such that $v_j(\pi(T')) \leq v_j(\tilde{\pi})$. Hence, any coalition T' different from $T(i)$ will be rejected.

If now player i receives an offer $T(i)$, any deviation will lead to the formation of the coalition $T(i)$, since any different offer by player i will be rejected by some player.

Finally, suppose that player i receives an offer $T' \neq T(i)$. If $v_i(\pi(T')) \leq v_i(\tilde{\pi})$, she cannot benefit from accepting the offer. If all other members of T' accept the offer, the coalition T' is formed and player i obtains a payoff $v_i(\pi(T')$, whereas, by rejecting the offer, player i obtains the payoff $v_i(\tilde{\pi})$. If $v_i(\pi(T')) > v_i(\tilde{\pi})$, player i should accept the offer, since her rejection would lead to the formation of the structure $\tilde{\pi}$, whereas her acceptance may either secure the formation of $\pi(T')$, if no player following player i rejects the offer T', or yield the formation of $T(i)$, if some player following player i rejects the offer T'.

Since player i has no incentive to deviate from her strategy $\tilde{\sigma}_i$, the strategy profile $\tilde{\sigma}$ forms a subgame perfect equilibrium of the game $\Gamma(v, \tilde{\rho})$. Furthermore, by construction, $\pi(\tilde{\sigma}) = \tilde{\pi}$. Hence, $CC(v) \subset SECS(v, \tilde{\rho})$. ■

In the statement of Proposition 3.2, I require the set of stationary perfect equilibria to be nonempty. This assumption is needed to show that, once a coalition structure is formed off the equilibrium path, the game still admits a stationary perfect equilibrium. The following example shows that the assumption cannot be relaxed.

The game of Example 3.3 admits a unique core stable structure, the grand coalition which Pareto dominates any other coalition structure. However, the subgame following the formation of the coalition $\{a\}$ is identical to the game in Example 2.6 and does not admit any stationary perfect equilibrium.

EXAMPLE 3.3 $N = \{a, b, c, d\}$.

π	$v_a(\pi)$	$v_b(\pi)$	$v_c(\pi)$	$v_d(\pi)$
$abcd$	5	5	5	5
$a\|bc\|d$	1	3	2	1
$a\|b\|cd$	1	1	3	2
$a\|bd\|c$	1	2	1	3
Others	1	1	1	1

The difficulty illustrated by Example 3.3 can be alleviated by assuming that, in addition to the valuation v, all restrictions of the valuation to subsets of the players admit a core stable structure.[13] Since payoffs depend on the whole coalition structure, the restriction of the valuation v to a subset K of the players must entail a description of the partition formed by the external players.

The *restriction* of the valuation v to a subset K of the players relative to the coalition structure $\pi_{N\backslash K}$ is defined as follows. $v(K, \pi_{N\backslash K})$: $\Pi_K \rightarrow \Re^{|K|}$ where $v(K, \pi_{N\backslash K})_{\mathrm{i}}$ $(\pi_K) = v_i(\pi_K \cup \pi_{N\backslash K})$.

Lemma 3.4. *Let v be a valuation such that $CC(v) \neq \emptyset$, and, for any restriction v' of v, $CC(v') \neq \emptyset$. Then, for any rule of order ρ, $SECS(v, \rho) \neq \emptyset$.*

Proof. Let ρ be a fixed rule of order. I construct a stationary perfect equilibrium strategy profile. For any restriction v' of v to a subset K of the players, relative to the coalition structure $\pi_{N\backslash K}$, pick a core stable structure. This core stable structure is denoted by $CS(\pi_{N\backslash K})$, and, for any player i in K, $T(i, \pi_{N\backslash K})$ denotes the coalition player i belongs to in $CS(\pi_{N\backslash K})$.

Construct a stationary strategy profile σ as follows.

$$\sigma_i\,(N\backslash K, \pi_{N\backslash K}, \emptyset) = T(i, \pi_{N\backslash K})$$

$$\sigma_i(N\backslash K, \pi_{N\backslash K}, T(i, \pi_{N\backslash K})) = \text{Yes}$$

$$\sigma_i(N\backslash K, \pi_{N\backslash K}, T') = \text{Yes} \quad \text{if } v_i(CS(\pi_{N\backslash K} \cup T')) > v_i(CS(\pi_{N\backslash K}))$$

$$\sigma_i(N\backslash K, \pi_{N\backslash K}, T') = \text{No} \quad \text{if } v_i(CS(\pi_{N\backslash K} \cup T')) \leq v_i(CS(\pi_{N\backslash K}))$$

To show that σ forms a subgame perfect equilibrium, consider all possible deviations for player i.

If player i proposes a coalition $T' \neq T(i, \pi_{N\backslash K})$, one of the members of T' will reject the offer, since $CS(\pi_{N\backslash K})$ is a core stable structure. Hence, player i cannot benefit from announcing a coalition different from $T(i, \pi_{N\backslash K})$. Similarly, by rejecting the offer $T(i, \pi_{N\backslash K})$, player i cannot obtain a higher payoff since the only coalition she can announce is the coalition $T(i, \pi_{N\backslash K})$.

Suppose now that player i receives an offer T' off the equilibrium path. By the same argument as in Proposition 3.2, she should accept the offer only if the payoff she receives in the final coalition structure is higher than the payoff she receives in $CS(\pi_{N\backslash K})$. The final coalition structure obtained after the formation of T', given the construction of the strategies, is the coalition structure $CS(\pi_{N\backslash K} \cup T')$. Hence, no deviation from the strategy σ_i can be profitable and the constructed strategy profile σ is a stationary perfect equilibrium. ■

Proposition 3.2 and Lemma 3.4 immediately lead to the following corollary.

[13] This requirement is very similar to the condition of total balancedness for games without externalities.

Corollary 3.5. *Let v be a valuation such that $CC(v) \neq \varnothing$, and, for all restrictions v' of v, $CC(v') \neq \varnothing$. Then, for any rule of order ρ, $CC(v) \subset SECS(v, \rho)$.*

Lemma 3.4 provides a sufficient condition for the existence of an equilibrium coalition structure. Corollary 3.5 shows that any core stable structure of a valuation v whose restrictions also admit core stable structures can be reached as the outcome of a stationary perfect equilibrium of the game of coalition formation. In the case of α stability, no such result can be expected. The following example shows that the sets of stationary equilibrium coalition structures and of α stable structures may be nonempty and disjoint.

EXAMPLE 3.6 $N = \{a, b, c, d, e\}$.

π	$v_a(\pi)$	$v_b(\pi)$	$v_c(\pi)$	$v_d(\pi)$	$v_e(\pi)$
$ab\vert cd\vert e$	4	4	3	3	1
$a\vert bc\vert d\vert e$	1	5	5	4	1
$ae\vert bc\vert d$	1	5	5	4	1
$a\vert bc\vert de$	1	1	1	5	5
$ac\vert b\vert de$	1	2	1	1	1
$a\vert b\vert c\vert de$	1	2	1	1	1
Others	1	1	1	1	1

The game admits three α stable structures $\{\{a\}, \{b, c\}, \{d\}, \{e\}\}$, $\{\{ae\}, \{bc\}, \{d\}\}$ and $\{\{a\}, \{bc\}, \{de\}\}$. To check that the coalition structure $\{\{a\}, \{b, c\}, \{d\}, \{e\}\}$ is α stable, observe that the only players who have an incentive to deviate are players d and e, who may want to form a coalition. However, their deviation is prevented by the formation of the coalition structure $\{\{a, c\}, \{b\}\}$ by the three other players. The coalition structure $\{\{ae\}, \{bc\}, \{d\}\}$ is α stable for the same reason. The structure $\{\{a\}, \{bc\}, \{de\}\}$ is α stable because the only two profitable deviations can be prevented by the external players. If players a and b form the coalition $\{a, b\}$, the three other players can react by forming the structure $\{\{c\}, \{d\}, \{e\}\}$, inducing a payoff of 1 for the two deviating players. If player b decides to break the coalition with player c, the four external players can form the coalition $\{a, b, c, d\}$ which yields a payoff of 1 for player b.

These three coalition structures are the only α stable structures of the game. The coalition structure $\{\{a, b\}, \{c, d\}, \{e\}\}$ is not α stable since players b, c and d can deviate and form the structure $\{\{b, c\}, \{d\}\}$ in which they are guaranteed to obtain higher payoffs. All other coalition structures are Pareto dominated by the coalition structure $\{\{a, b\}, \{c, d\}, \{e\}\}$ and hence are not α stable.

I now claim that the unique stationary equilibrium coalition structure of the game, independently of the rule of order ρ, is the coalition structure $\{\{a, b\}, \{c, d\}, \{e\}\}$. Two cases must be distinguished, one where ρ assigns as the first player a or b, one where c, d or e are chosen to start the game. If player a starts the game, player a should offer the formation of a coalition $\{a, b\}$. This offer will be accepted by player b, since, if player b were to form the coalition $\{b, c\}$, players d and e would form a coalition, inducing a payoff of 1 for player b. Given that players a and b have formed a coalition, player c should offer to form a coalition with player d, who will accept. Hence, in equilibrium, the coalition structure $\{\{a, b\}, \{c, d\}, \{e\}\}$ is formed. The same line of reasoning applies when player b starts the game.

If now player c starts the game, she should offer the formation of the coalition $\{c, d\}$, since the offer $\{b, c\}$ will be rejected by player b. This offer will be accepted by player d. In fact, player d has no incentive to form the coalition $\{d, e\}$ since this induces player b to form the coalition $\{b\}$. Once the coalition $\{c, d\}$ is formed, players a and b form the coalition $\{a, b\}$, yielding the coalition structure $\{\{a, b\}, \{c, d\}, \{e\}\}$. A similar line of reasoning applies to the cases where d and e start the game.

Example 3.6 is robust to small variations of the valuation. Hence there exists a class of valuations v, such that $SESC(v, \rho) \neq \emptyset$, $C\alpha(v) \neq \emptyset$ and $SECS(v, \rho) \cap C\alpha(v) = \emptyset$.

The absence of coincidence between the sequential game of coalition formation and the model of α stability stems from two countervailing forces in the definitions of deviations. On the one hand, deviations in the sequential model are *easier* to obtain, because the external players who have already formed a coalition cannot freely react to the deviation. This suggests that there may exist α stable structures which cannot be outcomes of subgame perfect equilibria of the game. In Example 3.6, the coalition structures $\{\{a\}, \{b, c\}, \{d\}, \{e\}\}$ and $\{\{ae\}, \{bc\}, \{d\}\}$ are not stationary equilibrium structures, because, once players a, b and c have left the game, players d and e can deviate and form the coalition $\{d, e\}$. Similarly, the coalition structure $\{\{a\}, \{bc\}, \{de\}\}$ cannot be obtained in a stationary perfect equilibrium, because b has an incentive to deviate after the coalition $\{d, e\}$ has been formed.

On the other hand, deviations in the sequential model are *harder* to obtain, because group deviations are not allowed, and players look forward to the final consequences of their deviations. Hence stationary equilibrium coalition structures are not necessarily α stable. In Example 3.6, the coalition structure $\{\{a, b\}, \{c, d\}, \{e\}\}$ is not α stable, because players b, c and d may deviate jointly and form the coalition structure $\{\{a\}, \{b, c\}, \{d\}, \{e\}\}$.

4. Sequential Formation of Coalitions in Symmetric Games

In this section, I analyze the formation of coalitions in the restricted class of symmetric games. Symmetric games are described by valuations where all players are ex ante identical. Hence, the payoffs received by the players only depend on coalition sizes and not on the identity of the coalition members.

Formally, let p denote a *permutation* of the players in N. For any coalition structure π of N, let $p\pi$ denote the coalition structure obtained by permuting the players according to p. A valuation v is *symmetric* if and only if $\forall i \in N$, $v_i(\pi) = v_{pi}(p\pi)$.

A *symmetric game* is a game described by a symmetric valuation. Observe that in symmetric games all members of a coalition receive the same payoff and payoffs only depend on the sizes of the coalitions. An important feature of symmetric games is that two coalition structures which only differ by the distribution on the players in the coalitions generate the same payoff distribution. This leads to the notion of *equivalence* of coalition structures in symmetric games.

Two coalition structures π and π' are called *equivalent* if there exists a permutation p of the players in N such that $\pi' = p\pi$. Two equivalent partitions are said to be equal up to a permutation of the players. The equivalence class of a coalition structure π is denoted by $eq(\pi)$. If the valuation v is symmetric, two equivalent partitions generate the same distribution of payoffs. Hence, in symmetric games, the study of coalitions can be restricted to the study of equivalence classes of partitions. An equivalence class of partitions can be identified with a list of coalition sizes, that is a sequence of positive integers adding up to n. I assume that the rule of order ρ is fixed, and let the players be indexed by the ordered set $I = 1, 2, \ldots, n$. This can be done without loss of generality, since any coalition structure emerging as an equilibrium of the game $\Gamma(v, \rho')$, for $\rho' \neq \rho$, is equivalent to a coalition structure generated by an equilibrium of the game $\Gamma(v, \rho)$. Since the rule of order ρ is fixed, the game Γ will only be indexed by the valuation v.

Since in a symmetric game, all players are ex ante identical, I restrict my attention to symmetric equilibria where all players adopt similar strategies. Formally, a strategy profile $\sigma = \{\sigma_i\}_{i \in N}$ is called *symmetric* if and only if (i) at any two states $s = (K, \pi_K, T)$, $s' = (K, \pi_K, T')$ with $|T| = |T'| \neq 0$, for any two players $i \in T$, $j \in T'$, $\sigma_i(s) = \sigma_j(s')$ and (ii) at any state $s = (K, \pi_K, \emptyset)$, for any two players i, $j \notin K |\sigma_i(s)| = |\sigma_j(s)|$. In words, a strategy profile is symmetric if, at any state, all responders adopt the same strategy and all proposers announce coalitions of the same size. The set of coalition structures supported by symmetric stationary perfect equilibria is denoted $SSECS(v)$.

I first show that, in a symmetric game, any symmetric stationary perfect equilibrium coalition structure can be reached as the outcome of a finite game of choice of coalition sizes. Furthermore, under a simple condition proposed by Ray and Vohra (1995), any equilibrium outcome of the game of choice of coalition sizes can be obtained as a symmetric stationary equilibrium coalition structure of the sequential game of coalition formation. Using this equivalence, I derive a sufficient condition under which a symmetric game admits a symmetric stationary equilibrium coalition structure and prove that this structure is generically unique.

The game of choice of coalition sizes $\Delta(v)$ is described as follows. Player 1 starts the game and chooses an integer k_1 in the interval $[1, n]$. Player $k_1 + 1$ then moves and chooses an integer k_2 in the set $[1, n - k_1]$. Player $k_1 + k_2 + 1$ chooses at the next stage an integer k_3 in the set $[1, n - k_1 - k_2]$. The game continues until the sequence of integers $(k_1, k_2, \ldots, k_j, \ldots, k_J)$ satisfies $\sum k_j = n$. The game for three players is depicted in Figure 3.

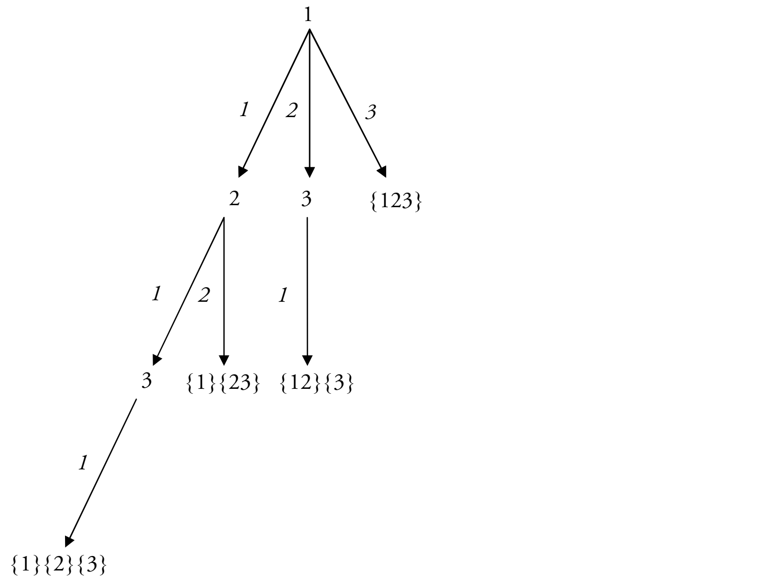


Figure 3. The game Δ

A strategy τ_i for player i in the game $\Delta(v)$ is a mapping from the set Π_{i-1} to the set of integers in the interval $[1, n - i - 1]$. In words, for any coalition structure π_{i-1} of the first $i - 1$ players, player i chooses a coalition size $\tau_I(\pi_{i-1})$. All players need not be called to announce coalition sizes in the game. Observe, however, that, for any strategy profile τ, a single coalition structure $\pi(\tau)$ is formed. The payoffs received by the players are then given by $v_i(\pi(\tau))$.

A strategy profile τ^* is a *subgame perfect equilibrium* if and only if for all players i, for all coalition structures π_{i-1} in Π_{i-1} and for all strategies τ_i, $v_i(\pi(\tau_i^*(\pi_{i-1}), \tau_{-i}^*)) \geq v_i(\pi(\tau_i(\pi_{i-1}), \tau_{-i}^*))$. As before, a coalition structure generated by a subgame perfect equilibrium τ^* is called an *equilibrium coalition structure* of the game $\Delta(v)$. The set of equilibrium structures of $\Delta(v)$ is denoted $ECS'(v)$.

Lemma 4.1. *For any symmetric valuation v, $ECS'(v) \neq \emptyset$.*

Proof. The game $\Delta(v)$ is a finite game of perfect information with perfect recall. Hence it admits a subgame perfect equilibrium in pure strategies. ■

In the next proposition, I show that any symmetric stationary equilibrium coalition structure of the game $\Gamma(v)$ can be reached as an equilibirum coalition structure of the game $\Delta(v)$, up to a permutation of the players.

Proposition 4.2. *For any coalition structure π in $SSECS(v)$ there exists a coalition structure π' equivalent to π such that $\pi' \in ECS'(v)$.*

Proof. Let σ be the symmetric stationary perfect equilibrium of the game $\Gamma(v)$ supporting the coalition structure π. I first show that the strategy profile σ cannot involve any delay and that all offers prescribed by σ are accepted. Suppose to the contrary that some player i rejects an offer T with $|T| \geq 2$ at some state $s = (K, \pi_K, T)$. Since the strategy profile σ is symmetric, $|\sigma_i(K, \pi_K, \emptyset)| = |T|$ and for all players $j \in \sigma_i(K, \pi_K, \emptyset)$, we have $\sigma_j(K, \pi_K, \sigma_i(K, \pi_K, \emptyset)) = \sigma_i(K, \pi_K, T) = \text{No}$. Hence offers are continuously rejected and the play of the game is infinite yielding a payoff of 0 to player i. Since however $\min_{\pi \supset \{\{i\}\}} v_i(\pi) > 0$, player i has an incentive to deviate and leave the game. This shows that, at a symmetric equilibrium σ, all offers are accepted. Hence the strategy σ can be described by a list of offers made by the players at all states where they are proposers.

As a second step, I show that we can assume without loss of generality that, at any two equivalent states $s = (K, \pi_K, \emptyset)$ and $s' = (K', \pi'_K, \emptyset)$ where $|K| = |K'|$ and the two coalition structures π_K and $\pi_{K'}$ are equivalent, $|\sigma_i(s)| = |\sigma_i(s')|$ to see this first reorder the players according to a rule of order $\hat{\rho}$ consistent with the order in which the coalition structure π is formed. Now, for any set K with $i \notin K$, let $\hat{K}$ denote the first $\hat{\rho}$-elements in $N \setminus \{i\}$ and, for any partition π_K of K , let $\hat{\pi}_K$ denote the equivalent partition of $\hat{K}$ Construct then the strategy $\hat{\sigma}_i$ as follows. At any state $s = (K, \pi_K, \emptyset)$, let $\hat{\sigma}_i(s)$ be a subset of $N \setminus K$ containing i such that $|\hat{\sigma}_i(K, \pi_K, \emptyset)| = |\sigma(\hat{K}, \hat{\pi}_K, \emptyset)|$. In words, I select for any state $s = (K, \pi_K, \emptyset)$ a particular representative of the equivalence class $eq(\pi_K)$ and $\hat{\sigma}_i$ assigns the action chosen for this representative state to the entire equivalence class. Clearly, the strategy $\hat{\sigma}$ satisfies the condition that sets of the same cardinality are chosen at two equivalent states. Furthermore, given the particular order $\hat{\rho}$ chosen, $\pi(\hat{\sigma}) = \pi(\sigma)$. It remains to show that $\hat{\sigma}$ forms a subgame perfect equilibrium of the game $\Gamma(v)$. To see this, consider a state $s = (K, \pi_K, \emptyset)$ and note that, since the strategy $\hat{\sigma}$ is played, any action of player i induces a unique partition of the set $N \setminus K$. Now suppose by contradiction that $\hat{\sigma}_i$ is not an optimal choice, i.e. that there exists a strategy $\tilde{\sigma}_i$ inducing a partition $\tilde{\pi}_{N \setminus K}$ such that $v_i(\pi_K \cup \tilde{\pi}_{N \setminus K}) > v_i(\pi_K \cup \hat{\pi}_{N \setminus K})$ where $\hat{\pi}_{N \setminus K}$ is the coalition induced by $\hat{\sigma}_i$. Next consider a permutation $\hat{p}$ of the players such that $\hat{p}K = \hat{K}$. Since the game is symmetric, $v_i(p(\pi_K \cup \tilde{\pi}_{N\setminus K})) = v_i(\pi_K \cup \tilde{\pi}_{N\setminus K}) > v_i(\pi_k \cup \hat{\pi}_{N/K}) = v_i(p(\pi_k \cup \hat{\pi}_{N/K}))$, contradicting the fact that σ_i is an optimal choice at $(\hat{K}, \hat{\pi}_K, \emptyset)$.

Since we may assume, by the preceding step, that the strategy σ assigns sets of the same cardinality at any two equivalent states, we are ready to construct a strategy profile τ in the game $\Delta(v)$ as follows. For any player i and any coalition structure π_{i-1} of the preceding players, let $\tau_i(\pi_{i-1}) = |\sigma_i(K, \pi_K, \emptyset)|$. To show that τ forms a subgame perfect equilibrium of the game $\Delta(v)$, suppose by contradiction that player i has a profitable deviation $\tau_i' \neq \tau_i$ after the coalition structure π_{i-1} is formed. I claim that this implies that player i has a profitable deviation from σ_i in the game $\Gamma(v)$. To see this, suppose that a coalition structure π_K equivalent to π_{i-1} has been formed and let player i reject any offer T such that $|T| \neq \tau_i'$ and propose the formation of a coalition T' of size τ_i'. Since τ_i' is a profitable deviation in the game $\Delta(v)$ and letting π' denote the coalition structure induced by the choice τ_i', we must have $v_i(\pi') > v_i(\pi)$. Now, by symmetry, for all players j in T', $v_j(\pi') = v_i(\pi') > v_i(\pi) = v_j(\pi')$, so that player i's offer is accepted. ■

While Proposition 4.2 guarantees that any symmetric equilibrium can be obtained as an equilibrium outcome of the game of choice of coalition sizes, it does not imply that the equilibrium coalition structures of the game Δ form symmetric stationary equilibrium outcomes of the sequential game of coalition formation. In fact, as noted by Ray and Vohra (1995), a stronger condition is needed for this assertion to hold : the coalitions formed in the game Δ must have the property that the players' payoffs are decreasing in the order in which coalitions are formed.

Proposition 4.3 *(Ray and Vohra, 1995). Let π be an equilibrium coalition structure of the game $\Delta(v)$ with the property that players' payoffs are decreasing in the order in which coalitions are formed. Then there exists a coalition structure π' equivalent to π such that $\pi' \in SSECS(v)$.*

Proof. Let τ be the subgame perfect equilibrium supporting τ. Define a strategy σ_i for player i in the game $\Gamma(v)$ as follows. At any state $s = (K, \pi_K, \emptyset)$ let player i announce a subset T of $N\setminus K$ with $|T| = \tau_j(\pi_{j-1})$ for the coalition structure π_{j-1} equivalent to π_K. At any state $s = (K, \pi_K, T)$ with $T \neq \emptyset$, let $\sigma_i(s) = \text{Yes}$ if $|T| = \tau_j(\pi_{j-1})$ and $\sigma_i(s) = \text{No}$ otherwise. This strategy profile is symmetric and yields a coalition structure $\pi(\sigma)$ equivalent to π. It remains to show that it forms a stationary perfect equilibrium of the game $\Gamma(v)$. First consider player i's possible deviation at a state $s = (K, \pi_K, \emptyset)$ when it is her turn to make an offer. If she makes any offer T' such that $|T'| \neq \tau_j(\pi_{j-1})$ and $|T'| \geq 2$, her offer will be rejected. Hence player i will belong to a coalition formed later in the game and, by assumption, her payoff is lower than the one she obtains in coalition T. By the same reasoning, player i has no incentive to reject an offer T where $|T| = \tau_j(\pi_{j-1})$. Finally, consider player i 's response to an offer T' with $|T'| \neq \tau_j(\pi_{j-1})$. By rejecting the proposal and offering to form a coalition T of size $|T| = \tau_j(\pi_{j-1})$, she can secure the formation of the coalition structure π. Since τ is a subgame perfect equilibrium of the game of choice of coalition sizes, $v_i(\pi) \geq v_i(\pi_K \cup \tilde{\pi}_{N\backslash K})$ for any other coalition structure $\tilde{\pi}_{N\backslash K}$ induced by the formation of a coalition T' at state $s = (K, \pi_K, \emptyset)$. Hence no player has any incentive to deviate from the strategy prescribed by σ. ■

Propositions 4.2 and 4.3 provide a sufficient condition on the underlying valuation v for the equivalence between the symmetric stationary perfect equilibrum outcomes of the sequential game of coalition formation and the subgame perfect equilibrium outcomes of the game of choice of coalition sizes. This result is formally stated in the next corollary.

Corollary 4.4. *Suppose that, in the game $\Delta(v)$, players' payoffs are decreasing in the order in which coalitions are formed. Then, for any coalition structure π in $SSECS(v)$ and any coalition structure π' in $ECS'(v)$, $eq(\pi) = eq(\pi')$.*

Hence, under a simple condition, the game of choice of coalition sizes provides an easy method for the construction of equilibrium coalition structures in symmetric games. The exact nature of the restriction that players' payoffs are decreasing in the order in which coalitions are formed is difficult to interpret. Ray and Vohra (1995) provide an example where the condition is violated and the subgame perfect equilibrium outcome of the game of choice of coalition sizes does not form a symmetric stationary perfect equilibrium of the sequential game. However, in most economic applications of coalitions with externalities, including the formation of cartels and of coalitions in majority games discussed in this paper, this condition is satisfied. The equivalence result of Corollary 4.4 can now be used to establish several important properties of equilibrium coalition structures in symmetric games.

Corollary 4.5. *Let v be a symmetric valuation such that, in the game $\Delta(v)$, players' payoffs are decreasing in the order in which coalitions are formed. Then $SECS(v) \neq \emptyset$.*

Proof. Follows from Lemma 4.1 and Corollary 4.4. ■

Corollary 4.4 also leads to a simple sufficient condition for the uniqueness of symmetric stationary equilibrium coalition structures in symmetric games. A valuation v is called *strict* if, for any player i, and for any two different partitions π and π', $v_i(\pi) \neq v_i(\pi')$. In a game described by a strict valuation, every agent receives different payoffs in different coalition structures. The next proposition shows that the strictness condition is sufficient to guarantee the uniqueness of the equilibrium coalition structure in the game $\Delta(v)$.

Proposition 4.6. *Let v be a strict symmetric valuation. Then the game $\Delta(v)$ has a unique equilibrium coalition structure.*

Proof. The proof is by induction on the number n of players. If $n = 1$, the game $\Delta(v)$ has a unique subgame perfect equilibrium. Suppose now that, for any $n' < n$, the game admits a unique subgame perfect equilibrium, and consider the first player's choices in a game with n players. For any choice of an integer k, the continuation game has less than n players, and thus admits a unique subgame perfect equilibrium $\tau^*(\{k\})$. Since the valuation is strict, there exists a unique k^*, such that

$$v_1(\{k^*\} \cup \pi(\tau^*(\{k^*\}) > v_1(\{k\} \cup \pi(\tau^*(\{k\}) \qquad \forall k \neq k^*.$$

Hence the n player game admits a unique subgame perfect equilibrium. ■

Proposition 4.6 implies that, if the valuation is strict, all equilibrium coalition structures of the game r(v) are equivalent. Hence I obtain the following corollary.

Corollary 4.7. *Let v be a strict symmetric valuation such that, in the game $\Delta(v)$, players' payoffs are decreasing in the order in which coalitions are formed. Then the game $\Gamma(v)$ has a unique symmetric stationary equilibrium coalition structure, up to a permutation of the players.*

5. Applications

In this section, I apply the sequential model of coalition formation to two particular symmetric situations. I first analyze the formation of cartels in a symmetric Cournot oligopoly. The second application is based on Hart and Kurz (1984)'s study of endogenous coalition formation in symmetric majority games. In both applications, I derive the subgame perfect equilibrium of the game of choice of coalition sizes. It is straightforward to check that players' payoffs are decreasing in the order in which coalitions are formed, so that the equivalence result of Corollary 4.4 can be applied.

5.1 Cartels in a symmetric Cournot oligopoly

It has long been noted that the formation of cartels in oligopolies involves a fundamental instability (See Stigler, 1968), since, once a cartel has been formed, members of the cartel obtain a lower profit than outsiders, and hence have an incentive to leave the cartel. Salant et al. (1983) analyze this instability in a simple symmetric Cournot oligopoly with linear demand and homogeneous goods, and show that there exists a minimal profitable size of the cartel which is never lower than four fifths of the members of the industry. This cartel is however (intuitively) unstable since members of the cartel would prefer to stay out and let the other firms form a cartel. In the sequential model analyzed here, firms have the power to commit to stay out of the cartel. Hence, the unique equilibrium coalition structure predicts that firms choose to remain outside of the cartel, until the remaining firms form the cartel of minimal profitable size.

More precisely, consider a Cournot oligopoly where firms face a linear inverse demand curve, $D = \alpha - \sum q_i$, where q_i is the quantity produced by each firm i. All firms are assumed to have a constant marginal cost of λ. Suppose that K cartels have formed on the market, and that the structure of cartels is given by $\pi = \{T_1, T_2,..., T_k,..., T_K\}$. Straightforward computations show that, in equilibrium, each cartel will produce $q_i^*(\pi) = (\alpha - \lambda)/(K + 1)$.[14] Hence, firm i in the cartel $T(i)$ of size $t(i)$ obtains a payoff of

$$P_i^*(\pi) = \frac{(\alpha - \lambda)^2}{t(i)(K+1)^2}$$

The problem of cartel formation can thus be summarized by the valuation defined by $v_i(\pi) = P_i^*(\pi)$.

Proposition 5.1. *Any equilibrium of the game of cartel formation is characterized by* $\pi^* = (T_I^* \{j\}_{j \notin T_I^*})$ *where* t^* *is the first integer following* $(2n+3-\sqrt{4n+5})/2$ *(If* $\sqrt{4n+5}$ *is an integer,* t^* *can take on the two values* $(2n+3-\sqrt{4n+5})/2$ *and* $(2n+5-\sqrt{4n+5})/2.$*)*

Proof. See the Appendix. ■

5.2. Coalitions in symmetric majority games

In their study of endogenous coalition formation, Hart and Kurz (1983) advocate a two-stage approach, where players evaluate their payoffs, in any coalition structure, according to a fixed rule (Owen, 1977)'s extension of the Shapley Value to games with coalition structures), and play a game of coalition formation using the value as their expected payoff. Owen (1977)'s value differs from Aumann and Drèze (1974)'s value in assuming that players bargain over the worth of the grand coalition, as opposed to the worth of the coalition they belong to in the coalition structure. The formation of a coalition is thus interpreted as a way for the players to modify the environment in which they bargain over the worth of the grand coalition.[15]

[14] It is important to note that the equilibrium quantity produced by each cartel only depends on the number of cartels on the market.

[15] The axiomatic derivation of the two different values are given in Aumann and Drèze (1974) and Hart and Kurz (1983). The differences are thoroughly discussed in Hart and Kurz (1983).

Owen (1977)'s value is computed, for any game in coalitional function form w, any coalition structure π and any player i as

$$\varphi_i(w, \pi) = E(w(\mathcal{P} \cup i) - w(\mathcal{P})),$$

where the expectation is taken over any random order which is consistent with the coalition structure π (i.e. ranks consecutively members of any coalition in the coalition structure) and $\mathcal{P}$ is the set of predecessors of i according to the random order.

Hart and Kurz (1984) apply Owen (1977)'s value to analyze the formation of coalitions in different types of games in coalitional function form. We consider here only symmetric majority games.

Definition 5.2. A symmetric majority game $M(n, m)$ is defined as follows. The number n is the total number of players, and the integer m (the majority) is any integer in the interval $[(n + 1)/2, n]$. The coalitional function is given by

- $w(T) = 0$ if $t < m$
- $w(T) = 1$ for $t \geq m$,

where T is any coalition, and t denotes the cardinality of coalition T.

To compute the Owen value in the symmetric majority game $M(n, m)$, let me consider a coalition structure π containing K coalitions, $\pi = \{T_1, T_2,..., T_k,... T_K\}$. The total number of random orders consistent with the coalition structure π is $K!t_1!t_2!...t_k!...t_K!$, where t_k denotes the number of elements of the coalition T_K. It is then clear that for the incremental value of player i to be positive, it must be that player i is ordered at position m in the random order. Denoting by $T(i)$ the coalition player i belongs to and letting $\omega_i(\pi)$ denote the number of orderings of the coalitions in π such that a member of the coalition $T(i)$ is at position m, I obtain the following simple expression for the Owen value

$$\phi_i(\pi) = \frac{\omega_i(\pi)}{t(i)K!}.$$

Hence I can now define the valuation $v_i(\pi) = \varphi_i(\pi)$. The characterization of the equilibrium coalition structures is made difficult by the lack of structure of the function $\omega_i(\pi)$. In the absence of general characterization results, Table I describes the equilibrium coalition structures of any symmetric majority game with $n \leq 10$.[16]

[16] The computations leading to the characterization of the coalition structures are not reproduced here and are available from the author.

TABLE I. Coalition structures in symmetric majority games

		n = 3		
m = 2	*m* = 3			
ab|c	*a|b|c*			
	abc			
		n = 4		
m = 3	*m* = 4			
abc|d	*a|b|c|d*			
	abcd			
		n = 5		
m = 3	*m* = 4	*m* = 5		
abc|d|e	*ab|cd|e*	*a|b|c|d|e*		
abc|de	*abcd|e*	*abcde*		
		n = 6		
m = 4	*m* = 5	*m* = 6		
abcd|e|f	*abc|de|f*	*a|b|c|d|e|f*		
abcd|ef		*abcdef*		
		n = 7		
m = 4	*m* = 5	*m* = 6	*m* = 7	
abcd|e|f|g	*abcde|f|g*	*abcdef|g*	*a|b|c|d|e|f|g*	
abcd|ef|g	*abcde|fg*	*ab|cd|ef|g*	*abcdefg*	
abcd|efg		*abc|def|g*		
		n = 8		
m = 5	*m* = 6	*m* = 7	*m* = 8	
abcde|f|g|h	*abcdef|g|h*	*abcdefg|h*	*a|b|c|d|e|f|g|h*	
abcde|fg|h	*abcdef|gh*		*abcdefgh*	
abcde|fgh	*abc|def|g|h*			
	abc|def|gh			
		n = 9		
m = 5	*m* = 6	*m* = 7	*m* = 8	*m* = 9
abcde|f|g|h|i	*abcdef|g|h|i*	*abcd|efg|hi*	*ab|cd|ef|gh|i*	*a|b|c|d|e|f|g|h|i*
abcde|fg|h|i	*abcdef|gh|i*	*abcd|efg|h|i*	*abcdefgh|i*	*abcdefghi*
abcde|fg|hi	*abcdef|ghi*			
abcde|fgh|i				
abcde|fghi				
		n = *10*		
m = 6	*m* = 7	*m* = 8	*m* = 9	*m* = *10*
abcdef|g|h|i | j	*abcdefg|h|i | j*	*abcdefgh|i | j*	*abcdefghi | j*	*a|b|c|d|e|f|g|h|i|j*
abcdef|gh|i | j	*abcdefg|hi | j*	*abcdefgh|ij*		*abcdefghij*
abcdef|gh|i j	*abcdefg|hi j*	*abcd|efg|h|i | j*		
abcdef|ghi | j		*abcd|efgh|ij*		
abcdef|ghi j				

The equilibrium coalition structures of symmetric majority games are not easily interpreted. When the majority required to win (m) is small, it appears that the minimal winning coalition forms, members of the winning coalition all obtain $1/m$ and external members, who obtain 0, organize themselves freely. When the number of votes required to win increases, the share of any member of the winning coalition decreases and it may become profitable to form smaller coalitions. This effect explains why the minimal winning coalition does not necessarily form in the symmetric majority games $M(5,4)$, $M(6,5)$, $M(7,6)$, $M(8,6)$, $M(9,7)$, $M(9,8)$ and $M(10,8)$. However, if all votes are required to win, the only equilibrium coalition structures are the grand coalition and the coalition consisting of singletons. In fact, in that case, the probability to win is independent of the size of the coalition, and players should always try to form the smallest coalitions. Hence, the only possible equilibrium coalition structures are the coalitions consisting of singletons and the grand coalition which yield the same payoff of $1/n$ to all players. Finally, it should be noted that Hart and Kurz (1984) observed that the majority game $M(10,8)$ has no α stable coalition structure. However, in my framework, an equilibrium coalition structure exists for this game.

6. Conclusions

In this paper, I analyze a sequential noncooperative game of coalition forma- tion when the rule of payoff division is fixed and payoffs depend on the whole coalition structure. The extensive form of the game is closely related to the extensive forms proposed by Selten (1981), Chatterjee et al. (1993) and Moldovanu (1992) for games of coalitional bargaining. I show that any core stable structure can be obtained as the outcome of a stationary perfect equilibrium, provided that the set of stationary perfect equilibria is nonempty. I analyze games described by symmetric valuations and provide a condition under which, when all the players are identical ex ante, the game admits a symmetric equilibrium coalition structure which is generically unique up to a permutation of the players. I also provide examples to show that stationary perfect equilibria may fail to exist in general valuations and that the noncooperative approach followed here is unrelated to standard cooperative game-theoretic solution concepts.

The determination of the equilibrium coalition structure in the sequential game of coalition formation is driven by two basic features of the extensive form. First, the exogenous rule of order imposes a fixed order of moves by players in the game. Depending on the valuation, players may have an advantage in moving first, second or in any other position in the game. The rule of order thus creates an asymmetry among players which is determined outside the game. An important direction for future research is to eliminate this asymmetry and to explore conditions under which the equilibrium of the extensive form game is independent of the rule of order. This line of research has been pursued by Moldovanu and Winter (1995) in the context of games of coalitional bargaining. They show that order independent equilibria only exist when the underlying game in characteristic function form, as well as all its restrictions, have nonempty cores.

The second important feature of the extensive form is the commitment power of the players. I assume that, by accepting the offer to join a coalition, players are bound to remain in that coalition whatever coalition structure the other players may form. This implies that coalitions are formed one after another and that coalitions may not compete to attract members. In fact, this sequential structure of the process of coalition formation is the feature of the extensive form which guarantees the existence of an equilibrium. Extensive form games where players do not commit to stay in a coalition can easily be constructed. The existence and characterization of equilibria in these games constitutes a difficult but important area for future research.

Finally, the model analyzed in this paper assumes that the coalitional worth is divided according to a fixed sharing rule. While this approach greatly simplifies the analysis, it clearly restricts the applicability of the model. The study of extensive form procedures allowing players to bargain over the worth of coalitions seems to me to be the foremost topic for future research.

Appendix: Proof of Proposition 5.1

The proof consists in three steps. In the first two steps, I explicitly construct the stationary perfect equilibria of the game. Observe first that the only payoff- relevant part of any history of the game is the number of coalitions which have already been formed. To fix notations, let K be *the number of coalitions already formed*, and m be the number of remaining players in the game, after a given history.

Step 1. After a given history, suppose that K coalitions have been formed, and that m players remain in the game. Suppose furthermore that, if a coalition of size μ is formed, the remaining $m - \mu$ players remain isolated. Then the optimal choice of μ is given by:

$$\mu^* = 1 \qquad \text{if} \quad m \leq (K+1)^2$$

$$\mu^* = \mathrm{m} \qquad \text{if} \quad m \geq (K+1)^2$$

Given that the remaining $m - \mu$ players form singletons, the optimal number of players in a coalition, μ^*, solves:

$$\max F(\mu) = \frac{(\alpha - \lambda)^2}{\mu(K + m - \mu + 2)^2}$$

subject to $1 \leq \mu \leq m$.

The function $1/\mu(K + m - \mu + 2)^2$ is strictly decreasing for $1 \leq \mu \leq (K + m + 2)/3$, and strictly increasing for $K + m + 2)/3 \leq \mu \leq m$. Hence, the optimal choice μ^* is either 1 or m. Now,

$$F(1) = \frac{(\alpha - \lambda)^2}{(K + m + 1)^2}$$

$$F(m) = \frac{(\alpha - \lambda)^2}{m(K + 2)^2}$$

Solving the quadratic in m, I obtain:

$F(1) \leq F(m)$ if and only if $m \geq (K + 1)^2$.

Step 2. The game admits two stationary perfect equilibria, given by

Strategy 1.

If $m < (K + 1)^2$	choose $\mu = 1$
If $(K + 1)^2 \leq m < (K + 2)^2 + 1$	choose $\mu = m$
If $(K + 2)^2 + 1 \leq m$	choose $\mu = 1$

Strategy 2.

If $m \leq (K + 1)^2$	choose $\mu = 1$
If $(K + 1)^2 < m \leq (K + 2)^2 + 1$	choose $\mu = m$
If $(K + 2)^2 + 1 < m$	choose $\mu = 1$.

The two equilibria only differ in the rules chosen to break ties. In the first equilibrium, if a player is indifferent between forming a cartel of size m or forming a singleton, she chooses to form a cartel. In the second equilibrium, she chooses to remain isolated. In the remainder of the proof, I focus on strategy 1, and show that *given that ties are broken according to the rule that indifferent players choose to form coalitions*, strategy 1 is the unique stationary perfect equilibrium of the game.

The proof is by induction on the number of remaining players in the game. If $m = 2$, the player before last chooses whether to form a cartel of size 2 or to remain isolated, in which case the last player remains isolated as well. Since $K \geq 0$, $2 < (K + 2)^2 + 1$. Hence strategy 1 prescribes that a cartel is formed if and only if $2 \geq (K + 1)^2$, and by Step 1, this is the unique optimal strategy for the player before last.

Suppose now that, for any $m' < m$, strategy 1 is the unique equilibrium strategy. Consider the different possibilities with m players remaining in the game.

If $m < (K + 1)^2$, then $\forall m' < m$, $m' < (K + 2)^2$. Hence, whatever coalition the player forms, all subsequent players choose to remain isolated. Then, by Step 1, the unique optimal strategy is to choose to form a singleton.

If now $(K+1)^2 \leq m < (K+2)^2 + 1$, similarly, $\forall m' < m$, $m' < (K+2)^2$.

Hence, irrespective of the coalition formed by the player, the subsequent players choose to remain isolated and, by Step 1, since $m \geq (K+1)^2$, the player should choose to form a coalition of size m.

Finally, when $m \geq (K+2)^2 + 1$, different possibilities have to be considered. The player may either choose to form a coalition μ such that $(m - \mu) \geq (K+2)^2$, in which case the remaining players form a coalition, or a coalition μ such that $(m - \mu) < (K+2)^2$, in which case the remaining players choose to remain separate.

When the coalition size μ is such that $(m - \mu) < (K+2)^2$, the player's payoff is given by:

$$F(\mu) = \frac{(\alpha - \lambda)^2}{\mu(K + 2 + m - \mu)^2}\ .$$

From Step 1, since $m > (K+1)^2$, the optimal choice of coalition size is $\mu^* = m$. In the case where μ is chosen small enough, other players form a coalition later in the game. Given the specification of the strategy, after the formation of the coalition of size μ, a group of players will choose to remain separate, and the last players will form a single coalition. The number of players who choose to remain isolated, ν, is the unique integer satisfying:

$$(K + 2 + \nu)^2 \leq (m - \mu - \nu) < (\mathrm{K} + 3 + \nu)^2 + 1.$$

A simple computation shows that ν is the first integer following:

$$\nu^* = \frac{\sqrt{9 + 4(K + m - \mu)} - (2K + 5)}{2}\ .$$

Hence, the payoff to a player who chooses a coalition of size μ where $m - \mu \leq (K+2)^2$ is given by:

$$G(\mu) = \frac{(\alpha - \lambda)^2}{\mu(K + 3 + \nu^*)^2}\ ,$$

or

$$G(\mu) = \frac{(\alpha - \lambda)^2}{\mu(\sqrt{9 + 4(K + m - \mu)} + 1)^2}\ .$$

The optimal value μ^* is thus the minimum over the interval $[1, m - (K+2)^2]$ of the function

$$H(\mu) = \mu(\sqrt{9 + 4(K + m - \mu)} + 1)^2$$

Next consider the derivative H' of H,

$$H'(\mu) = (\sqrt{9+4(K+m-\mu)}+1)(\sqrt{9+4(K+m-\mu)}+1-\frac{4\mu}{(\sqrt{9+4(K+m-\mu)}}) .$$

A study of the sign of H' shows that the function H is increasing up to the value $\mu = (16K+16m+35+\sqrt{73+32K+32m})/32$, and decreasing thereafter.

Hence, the optimal choice of μ, μ^*, is either $\mu^* = 1$, or $\mu^* = m - (K+2)^2$. Now, a simple computation shows that the choice $\mu^* = m - (K+2)^2$ is dominated by $\mu^* = m$.

To complete this step, it suffices to show that $\mu^* = 1$ is the optimal choice, that is that $H(1) \leq m\ (K+2)^2$.

$$\begin{aligned} H(1) &= \tfrac{1}{4}(\sqrt{9+4(K+m-1)}+1)^2 \\ &= \tfrac{1}{2}(3+2K+2m+\sqrt{9+4(K+m-1)}) \\ &< \tfrac{1}{2}(3+2K+2m+9+4(K+m-1) \\ &< (3K+3m+4) \end{aligned}$$

Hence,

$$\begin{aligned} m(K+2)^2 - H(1) &> m(K+2)^2 - (3K+3m+4) \\ &> m(K^2+4K+1) - 3K - 4 \\ &> (K+2)^2(K^2+4K+1) - 3K - 4 \\ &> 0 \end{aligned}$$

Step 3. The coalition structure generated by the stationary perfect equilibria corresponding to strategies 1 and 2 is given by $\pi^* = (T_1^*\ \{\ j\ \}_{j\ \in T_1^*})$ where t_1^* is the first integer following $(2n+3-\sqrt{4n+5})/2$. (If $\sqrt{4n+5})$ is an integer, t_1^* can take on the two values $(2n+3-\sqrt{4n+5})/2$ and $(2n+5-\sqrt{4n+5})/2)$.

When $K = 0$, strategy 1 prescribes that the first player forms a singleton. In fact, singletons will continue to be formed as long as $m \geq (n - m + 2)^2 + 1$. The unique coalition formed will comprise t^* members, where t^* is the unique integer such that

$$(n - t^* + 1)^2 \leq t^* < (n - t^* + 2)^2 + 1.$$

References

Aumann, R. (1967), 'A survey of games without side payments', in M. Shubik (ed.), *Essays in Mathematical Economics*, Princeton, NJ: Princeton University Press, pp. 3–27.

Aumann, R. and J. Drèze (1974), 'Cooperative games with coalition structures', *International Journal of Game Theory* **3**, 217–237.

Aumann, R. and M. Maschler (1964), 'The bargaining set for cooperative games', in M. Dresher, L. Shapley and A. Tucker (eds.), *Advances in Game Theory*, Princeton, NJ: Princeton University Press, pp. 443–447

Aumann, R. and R. Myerson (1988), 'Endogenous formation of links between players and of coalitions: An application of the Shapley value', in A. Roth (ed.), *The Shapley Value: Essays in Honor of Lloyd Shapley*, Cambridge, UK: Cambridge University Press, pp. 175–191

Bloch, F. (1995), 'Endogenous structures of association in oligopolies', *RAND Journal of Economics* **26**, 537–556.

Chatterjee, K., B. Dutta, D. Ray and K. Sengupta (1993), 'A noncooperative theory of coalitional bargaining', *Review of Economic Studies* **60**, 463–477.

Fudenberg, D. and D. Levine (1983), 'Subgame perfect equilibria of finite and infinite horizon games', *Journal of Economic Theory* **31**, 251–268.

Greenberg, J. (1995), 'Coalition structures', in R. Aumann and S. Hart (eds.), *Handbook of Game Theory with Applications*, Vol. II, Amsterdam: North Holland, pp. 1305–1337.

Guesnerie, R. and C. Oddou (1981), 'Second best taxation as a game', *Journal of Economic Theory* **25**, 67–91.

Hart, S. and M. Kurz (1983), 'Endogenous formation of coalitions', *Econometrica* **51**, 1047–1064.

Hart, S. and M. Kurz (1984), 'Stable coalition structures' in M. Holler (ed.), *Coalitions and Collective Action*, Vienna: Physica Verlag, pp. 236–258.

Kurz, M. (1988), 'Coalitional value', in A. Roth (ed.), *The Shapley Value: Essays in Honor of Lloyd Shapley*, Cambridge, UK: Cambridge University Press, pp. 155–173.

Moldovanu, B. (1992), 'Coalition-proof Nash equilibria and the core in three-player games', *Games and Economic Behavior* **4**, 565–581.

Moldovanu, B. and E. Winter (1995), 'Order independent equilibria', *Games and Economic Behavior* **9**, 21–34.

Myerson, R. (1977), 'Graphs and cooperation in games', *Mathematics of Operations Research* **2**, 225–229.

Myerson, R. (1978), 'Threat equilibria and fair settlements in cooperative games', *Mathematics of Operations Research* **3**, 265–274.

Owen, G. (1977), 'Value of games with a priori unions', in R. Hein and O. Moeschlin (eds.), *Essays in Mathematical Economics and Game Theory*, New York: Springer-Verlag, pp. 76–88.

Ray, D. and R. Vohra (1995), 'Binding agreements and coalitional bargaining', mimeo, Department of Economics, Boston University and Brown University.

Rubinstein, A. (1982), 'Perfect equilibrium in a bargaining model', *Econometrica* **50**, 97–109.

Salant, S., S. Switzer and R. Reynolds (1983), 'Losses from horizontal mergers: The effects of an exogenous change in industry structure on Cournot-Nash equilibrium', *Quarterly Journal of Economics* **98**, 185–199.

Selten, R. (1981), 'A noncooperative model of characteristic function bargaining', in V. Bohm and H. Nachtkamp (eds.), *Essays in Game Theory and Mathematical Economics in Honor of Oskar Morgenstern*, Mannheim: Bibliographisches Institut Mannheim, pp. 131–151.

Shenoy, P. (1979), 'On coalition formation: A game theoretical approach', *International Journal of Game Theory* **8**, 133–164.

Shubik, M. (1982), *Game Theory in the Social Sciences*, Cambridge, MA: MIT Press.

Stigler, G. (1968), *The Organization of Industry*, Homewood, IL: Irwin.

Sutton, J. (1986), ‘Noncooperative bargaining theory: An introduction’, *Review of Economic Studies* **53**, 709–724.

Thrall, R. and W. Lucas (1963), ‘N-person games in partition function form’, *Naval Research Logistics Quarterly* **10**, 281–298.

Von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.

Winter, E. (1993), ‘Mechanism robustness in multilateral bargaining’, mimeo, Center for Rationality and Interactive Decision Theory, The Hebrew University of Jerusalem.

Yi, S.S. and H.S. Shin (1995), ‘Endogenous formation of coalitions. Part I: Theory’, mimeo, Department of Economics, Dartmouth College.

Pure Strategy Nash Equilibrium in a Group Formation Game with Positive Externalities

Hideo Konishi, Michel Le Breton and Shlomo Weber

Paper presented at the CTN Workshop 1997. Reprinted from *Games and Economic Behavior,* **21**(1–2), October 1997, 161–182, with permission from Elsevier.

This paper identifies a domain of payoff functions in no spillover noncooperative games with Positive externality which admit a pure strategy Nash equilibrium. Since in general a Nash equilibrium may fail to exist, in order to guarantee the existence of an equilibrium, we impose two additional assumptions, Anonymity and Order preservation. The proof of our main result is carried out by constructing, for a given game G, a potential function Ψ over the set of strategy profiles in such a way that the maximum of Ψ yields a Nash equilibrium in pure strategies of G.

1. Introduction

One can easily find a variety of examples to support the claim that our economic and social life is often conducted within the structure of *groups* of agents. Individual consumers are, in fact, households, and the individual producers are, in fact, firms which are coalitions of owners of different factors of production. Society produces its public goods within a complex web of federal, state and local jurisdictions, and the political life is conducted through the rather complicated structure of political parties.

We thank Charles Blackorby, Jim Dolmas, Peter Fishburn, Paolo Ghirardato, Marci Rossell, Suzanne Scotchmer, Tomoichi Shinotsuka, William Thomson, Takashi Ui, and participant s of the Economic Theory Meetings in Cephalonia and Milan, Game Theory Conference in Toronto, SETIT Conference in Dallas, the seminars at Caltech, CORE, CentER, TCER, Universities of Pennsylvania, Rochester, Toulouse, Venice, Vienna, and Kyoto, Osaka, Stanford, and Tezukayama Universities for their valuable comments and discussions. We are also grateful to the associate editor of this journal and the anonymous referee for their useful suggestions.

he reason for the existence of groups which usually contain more than one agent but less than the entire society lies in the conflict between increasing returns to scale provided by large groups on one hand and heterogeneity of agents' preferences on the other hand. Indeed, it is often the case that firms create joint research ventures rather than conducting R&D independently in order to extract the gains from cooperation and obtain access to a larger pool of resources. However, given the heterogeneity of agents' tastes, a decision-making process in large groups may lead to outcomes quite undesirable for some of its members. This observation supports the claim that, on many occasions a decentralized organization is superior to a large social structure. Instead of the grand coalition containing all agents in the economy, we often observe the emergence of coalition structures which consists of groups smaller than the entire society.

In this paper, we will analyze the issue of stability of endogenously formed group structures in *games with no spillovers*, where the *no spillover* condition means that for every group of players choosing the same strategy, the payoff of every member of this group is independent of choices made by players outside of the group. Given the complexity of the general problem, there is not much hope for a stability result which will hold for the entire class of 'group formation' games. We shall therefore focus on subclasses of these games which may yield stable group structures.

The goal of this paper is to study an interesting subclass of environments, satisfying not only the no spillover condition but also *positive externality* (*PE*), where increasing returns to the size of groups are reflected by the assumption that each player would enjoy a higher payoff from a given alternative in a larger group. One of the natural examples of such environments are those with 'network externalities' where the utility that a given user derives from the good depends upon the number of other users who are in the same 'network' as she. Consider, for example, the choice of word processors in a department . If there are many users of *Word*, it might be beneficial to 'join the crowd' and become a *Word* user, a decision to be welcomed by other *Word* users. Thus, this environment satisfies the positive externality condition. Moreover, if one of users of T^3 decides to switch to *Word Perfect*, the utility of the *Word* users would not be affected, thus yielding the no spillover condition. As Katz and Shapiro (1985) pointed out, there are several possible sources of positive consumption externalities.[1] It could be through a *direct* physical effect of the number of purchasers on the quality of the product, where the utility that a consumer derives from purchasing a telephone, for example, depends on the number of other households or businesses that have joined the telephone network. There may be *indirect* effects that give rise to consumption externalities as well. For example, an individual purchasing a personal computer is affected by the

[1] Liebowitz and Margolis (1994) referred to these circumstances as *network effects*.

number of other individuals or firms purchasing similar hardware because the amount and variety of software supplied for use with a given computer is usually an increasing function of the number of hardware units that have been sold.

Another example is the local public goods economies (Guesnerie and Oddou, 1981; Greenberg and Weber, 1986), where the members of every jurisdiction select a public good provision vector or a tax schedule, and each player is free to make her residential choice by comparing population composition and policy in each jurisdiction. Here the no spillover requirement simply means that a migration of an individual from jurisdiction *B* to jurisdiction *C* would not affect the utility of residents of jurisdiction *A*. Moreover, at least in small jurisdictions, cost benefits generated by sharing a burden of production of public goods among residents may outweigh congestion effects, in which case the cost of provision of a given level of public good declines with a number of residents in the jurisdiction. A possibility of the conflict between agglomeration and diversification arises here very naturally, since, for example, some residents could regard public education as quite important, while others might put a higher priority on police and fire protection services. The heterogeneity of agents' preferences for public services may explain an emergence of a large number of different jurisdictions in the society, thus *under lining* the importance of study of group formation.

In order to study different notions of stability of group structures we consider a game in normal form where each player chooses an alternative (strategy) from a common alternative set. The payoff of each player, obviously, depends on her chosen alternative and the set of players who made the same choice. We shall then identify a domain of payoff functions which yields the existence of a pure strategy Nash equilibrium when the number of players is finite. A Nash equilibrium is a relatively weak stability requirement which, in context of local public goods economy, represents the vector of individuals' residential choices that no one would benefit by moving to another jurisdiction. Since our primary interest is in the study of group structures generated by unambiguous pure strategy individuals' decisions (such as jurisdictional choice or selection of word processor), we consider in this paper Nash equilibria in pure strategies only. We shall also examine a notion of a *strong Nash equilibrium in pure strategies* when no *group* of individuals would benefit by jointly switching their alternatives. Since it is immune against not only individual but also any coalitional deviations, the notion of strong Nash equilibrium is, obviously, much more restrictive than that of a Nash equilibrium.[2]

[2] Although it is not our goal here, we may consider alternative stability notions of group structures by employing other solution concepts, e.g., coalition-proof Nash equilibrium (Bernheim et al., 1987), which makes use of only credible coalitional deviations. Since the set of strong Nash equilibria is a subset of the set of coalition-proof Nash equilibria and the latter is a subset of the set of Nash equilibria, some of our results may be used to examine the existence of a coalition-proof Nash equilibria.

First we consider the case where, as in many models with network externalities (Farre ll and Saloner, 1985, 1988; Tirole, 1988; Arthur, 1989), the set of pure strategies for each player consists of two alternatives. We show then that the *PE* assumption alone guarantees the existence not only of a Nash equilibrium but even a strong Nash equilibrium (Proposition 2.2). The result is quite strong since it yields existence of a stable group structure while using a very demanding notion of stability. However, Proposition 2.2 cannot be extended to the case where the number of alternatives is larger than two. Thus, we impose a number of additional conditions on players' payoff functions. The first is that of *Anonymity* (*AN*) which requires that each player's payoff depends only on the *number* of players who make the identical choices and is independent of the names of individuals in each group. *Order preservation (OP*), together with anonymity, implies that for every player i the ranking of any two alternatives a and b remains the same if we add (or delete) the same individual to (or from) the set of those who choose these two alternatives. Our main result (Proposition 4.1) states that, for any number of alternatives, the assumptions *PE*, *AN*, and *OP* guarantee the existence of a (pure strategy) Nash equilibrium of a no spillover game. To prove this result, we show first that there is a convenient utility representation of each individual's preferences. Using this representation, we then construct a *potential* function and demonstrate that its maximum gives rise to a (pure strategy) Nash equilibrium of our game. We show that Proposition 4.1 is tight in the sense that if we drop AN, a (pure strategy) Nash equilibrium might fail to exist even when the preferences of each player are single-peaked over the set of alternatives X (Example 4.4). We also point out that if we replace assumption *OP* by a weaker property *order invariance* (*OI*), which requires that for each player i the ranking over any two alternatives a and b remains the same as long as a and b are selected by the same number of players, then the set of (pure strategy) Nash equilibria might be empty even if individuals' preferences are single-peaked (Example 3.1) It turns out that if we drop *PE* even for one player, a (pure strategy) Nash equilibria might fail to exist (Example 4.5). Furthermore, we show that, unlike in Greenberg and Weber (1986, 1993), a strong Nash equilibrium does not, in general, exist under the same assumptions that guarantee the existence of a pure strategy Nash equilibrium (Example 4.6).

There are several papers which dealt with the existence of an equilibrium in local public good economy without congestion. First, Guesnerie and Oddou (1981) have derived the conditions on profiles of individuals' preferences which yield the existence of the core of the associated game in characteristic function form. Greenberg and Weber (1986) prove the existence of a strong Nash equilibrium in a game with linearly ordered players whose preferences satisfy a kind of *single crossing property* condition. Demange (1994) generalized their result to the case where the hierarchical structure of the game is represented by a *tree* rather than a straight line as in the Greenberg and Weber (1986) model.

It is important to mention that all assumptions in this paper, except in Section 5, are imposed directly on individual preferences rather than on profiles of individuals' preferences. In contrast, all aforementioned papers on local public goods impose various restrictions on profiles of preferences (a version of a single-crossing property in Greenberg and Weber, 1986, and 'intermediate' profiles of preferences in Demange, 1994). Given a specific economical or political environment, it would be rather easy to verify the applicability of our results by simply checking whether the preferences of every individual satisfy positive externality, anonymity, or other assumptions than to deal with intricate conditions on profiles of preferences. Moreover, our main proposition does not make any use of single-peaked preferences which are imposed in the vast majority of the studies on local public goods economies. Thus, our model is quite different from those examined in the aforementioned papers.

The paper most closely related to our model is that by Greenberg and Weber (1993), who proved the existence of a strong Nash equilibrium in the political party formation game. In this game each player has single-peaked preferences over a unidimensional set of alternatives, where the set of feasible alternatives of every party would expand with an increase in the number of its supporters. Thus, the players' utilities are not directly affected by the number of players in a group, although the set of players choosing a given alternative affects its feasibility. Thus, their condition is stronger than the *PE* assumption used in Greenberg and Weber (1993). In addition, we do not make use of single-peakedness in order to prove the existence of a (pure strategy) Nash equilibrium.

It is interesting to compare our results with the literature on congestion games. This class of games satisfies the *negative externality* assumption *NE* that represents decreasing returns to size: each player i is worse off if more players make an identical choice to that of i and thus join i's group. Recently, Milchtaich (1996), Konishi et al. (1997a) and Quint and Shubik (1994) independently proved the existence of a pure strategy Nash equilibrium in no spillover games that satisfy *NE* and *AN*. Konishi et al. (1997a) show that under the same conditions there exists even a strong Nash equilibrium. Proposition 4.1 of this paper implies that if *NE* is replaced by *PE*, much more stringent conditions are needed in order to guarantee the existence of a pure strategy Nash equilibrium. Moreover, Example 4.6 shows that even assumptions imposed in Proposition 4.1 do not, in general, yield the existence of a strong Nash equilibrium.

As we indicated above, the main purpose of this paper is to identify a class of the no spillover games with positive externality which admits the existence of a Nash equilibrium *without* imposing any restrictions on profiles of players' preferences. However, it is worth pointing out that the method of the proof of Proposition 4.1 can be used to obtain the existence of a Nash equilibrium for a more general class of preferences useful in many applications. The payoff functions in this class satisfy are separable in terms of individuals' preferences over alternatives and 'externality' effect. This effect is assumed to be common for all players and we show (Example 5.1) that this property is necessary to obtain the existence result. The paper is organized as follows: in the next section we present the model and state our first existence result for the case of two alternatives. In Section 3 we introduce additional assumptions and show that even *PE*, *AN*, and *OI* do not guarantee the existence of a pure strategy Nash equilibrium in no-spillover games. In Section 4 we prove our main existence result and show the tightness of our assumptions. In Section 5 we demonstrate the existence of a pure strategy Nash equilibrium under the restriction on the preferences' profiles and show that the common externality effect cannot be dispensed with. The utility representation result used in the proof of the main proposition is relegated to the Appendix.

2. The Model

Let X be a (finite or infinite) set of alternatives and N be a finite set of players. Each player i in N chooses an alternative x^i from the set X which is common to all players. The players' choices constitute a (vector representation of) strategy profile $\mathbf{x} = (x^1, x^2, ..., x^n)$. The set of all strategy profiles is, therefore, given by the product $X^N = X \times X \times \cdots \times X$. Each player $i \in N$ has a preference ordering over strategy profiles which is represented by utility function $U^i : X^N \rightarrow \Re$. The noncooperative game G is therefore represented by the triple (N, X, U) where $U = \{U^i\}_{i \in N}$ is the profile of players' preferences. The main purpose of the paper is to derive sufficient conditions which yield the existence of a pure strategy Nash equilibrium of game G. Since we exclusively deal with pure strategy Nash equilibria, no confusion will arise when we use simply a 'Nash equilibrium' instead of a 'pure strategy Nash equilibrium of the game G'.

Before introducing our assumptions, it is useful to observe that every strategy profile $\mathbf{x} = (x^1, x^2, ..., x^n)$ generates the partition $P(\mathbf{x})$ of the set of players over the alternative set according to their choices at $\mathbf{x}$. Since the set of alternatives X is common for all players, we may represent this partition as $P(\mathbf{x}) = (N_x(\mathbf{x}))_{x \in X}$, where for each $x \in X$ the set $N_x(\mathbf{x})$ denotes the set of those players who choose alternative x under the strategy profile $\mathbf{x}$. Obviously, the partition $P(\mathbf{x})$ is uniquely determined by the strategy profile $\mathbf{x}$. It is important to observe that the correspondence between the strategy profiles and the associated partitions is one-to-one, and from the given partition $P(\mathbf{x})$ one can reproduce the strategy profile $\mathbf{x}$.

We shall restrict our attention to *no-spillover* games, where for each player i her payoff is not affected by the players whose choices are different from her strategy x^i. That is, if the choice of player j, x^j is different from x^i, the payoff of i would not be affected if j switches her choice to any alternative x different from

x^i. Formally, for any strategy profiles $\mathbf{x}, \mathbf{y} \in X^N$ and an alternative $x \in X$ such that $N_x(\mathbf{x}) = N_y(\mathbf{y}) \neq \varnothing$, the payoff of every player $i \in N_x(\mathbf{x})$ is the same at $\mathbf{x}$ and $\mathbf{y}$, that is, $U^i(\mathbf{x}) = U^i(\mathbf{y})$. Note that the no-spillover games allow for the simple representation of the payoff function of player i when she chooses alternative x given the strategy profile $\mathbf{x}$:

$$U^i(\mathbf{x}) = u^i(x, N_x(\mathbf{x})).$$

As we mentioned above, some natural examples of environments without spillovers are given by local public good economies and location choice games. In the example of a word processor choice, the no-spillover condition implies that the utility of T^3 users would not be affected if some faculty member switches from Word Perfect to Word. In this paper we consider only the no-spillover games.

We first introduce the focus of this paper, *positive externality*, which requires that player *i*'s payoff increases if another player *j* who previously chose $x^j \neq x^i$ changes her strategy to x^i.

Assumption Positive Externality (PE). For any two players $i, j \in N$, for any subset of players $S \subset N$ with $i \in S$ and $j \notin S$ and alternative $x \in X$ we have $u^i(x, S) \leq u^i(x, S \cup \{j\})$.

Condition *PE* allows us to derive our first result. It states that in the case where the set of alternatives *X* consists of two elements, *PE* yields the existence of a Nash equilibrium of game *G*. The result is, in fact, even stronger as it yields the existence of a *strong* Nash equilibrium (Aumann, 1959). To recall:

Definition 2.1. A strategy profile $\mathbf{x} = (x^1, x^2, ..., x^n)$ is a ***strong Nash equilibrium of game*** *G* if there exists no subset *S* of *N* and a strategy $\bar{x}^j$ for each *j* in *S*, such that $U^j(\bar{\mathbf{x}}) > U^j(\mathbf{x})$ for all $j \in S$, where $\bar{\mathbf{x}}$ denotes the strategy profile which assigns $\bar{x}^j$ to every *j* in *S* and x^i to every player *i* who does not belong to *S*.

That is, a strong Nash equilibrium is immune to any group deviation. Thus, it is a much stronger equilibrium concept than a Nash equilibrium which allows only for individual deviations. Our first result is:[3]

Proposition 2.2 *Let* $|X| = 2$. *Then, under PE, every (no spillover) game G admits a strong Nash equilibrium.*[4]

[3] An alternative proof of this result could be derived by using Theorem 1 in Greenberg and Weber (1993).

[4] It is worthwhile to note a game whose set of alternatives consists of two elements trivially satisfies the no-spillover condition.

Proof. Let $X=\{a, b\}$. Consider first the strategy profile $\mathbf{x}_0$ which assigns the alternative a to every player in N. If it is a strong Nash equilibrium of game G, we are done. Suppose, therefore, that there exists a coalitional deviation that benefits all deviating players, that is, there is a coalition S such $u^i(b,S)>u^i(a,N)$ for all $i\in S$. Let S_1 be the maximal coalitional deviation (with respect to inclusion). By *PE*, S_1 contains any other possible coalitional deviations from $\mathbf{x}_0$. Consider the strategy profile $\mathbf{x}_1$ which assigns alternative b to all players in S_1 and alternative a to all other players, or, by using the partition representation of strategy profiles, $(N_a(\mathbf{x}_1),N_b(\mathbf{x}_1))=(N\setminus S_1,S_1)$. We shall show that:

Claim. A strategy profile x_1 is a strong Nash equilibrium if and only if there is no subset S_2 of the set $N\setminus S_1$ which can benefit all of its members by switching to alternative b; i.e., $u^i(b,S_1\cup S_2)>u^i(a,N\setminus S_1)$ for each $i\in S_2$.

Proof. Suppose that $\mathbf{x}_1$ is not a strong Nash equilibrium. Then there exists a deviation by coalition T which makes all members of T better off relatively to their payoff at $\mathbf{x}_1$ Since no subset of S_1 can make all its members better off by offering some of them alternative a, the set $\overline{T}=T\cap(N\setminus S_1)$ is nonempty.

Since $u^i(b,S_1)>u^i(a,N)$ for all $j\in S_1$, it follows that no member of $T\setminus\overline{T}$ can increase her payoff by switching back to a. Thus, every player in $\overline{T}$ is offered b by the coalitional deviation T. Hence, all players in $\overline{T}$ are better off by switching to b. ■

The claim implies that if $\mathbf{x}_1$ is not a strong Nash equilibrium, then there exists a subset S_2 of the set N \ S_1 which can benefit all of each members by switching to alternative b; i.e., $u^i(b,S_1\cup S_2)>u^i(a,N\setminus S_1)$ for each $i\in S_2$. Let S_2 be the maximal (with respect to inclusion) set satisfying this property. Consider then the strategy profile $\mathbf{x}_2$ whose partition representation is given by $(S_1\cup S_2,N\setminus(S_1\cup S_2))$. By using the arguments of the above claim repeatedly, we can show that the profile $\mathbf{x}_2$ is a strong Nash equilibrium if and only if there is no subset S_3 of the set $N\setminus(S_1\cup S_2)$ which can benefit all of each member by switching to alternative b, i.e., $u^i(b,S_1\cup S_2\cup S_3)>u^i(a,N\setminus(S_1\cup S_2))$ for each $i\in S_3$. We continue this procedure by picking the largest coalitional deviation from the set of players choosing a and adding them to the set of players choosing b. Since the set of players n is finite, the procedure will be terminated in a finite number of steps. That is there exists a number K such that there is no subset of $N\setminus(S_1\cup S_2\cup\cdots\cup S_K)$ which can benefit its members by switching from a to b. The repeated use of the arguments of the above claim would imply that the strategy profile $\mathbf{x}_K$ whose partition representation is given by $(N\setminus(S_1\cup S_2\cup\cdots\cup S_K),S_1\cup S_2\cup\cdots\cup S_K)$ is, indeed, a strong Nash equilibrium of the game G. ■

In the next section we shall provide an example that will illuminate the difficulties which do not allow us to extend Proposition 2.2 to the case with more than two alternatives.

3. Example

Before introducing the example which shows that a Nash equilibrium may fail to exist when the set of alternatives consists of more than two alternatives, we shall introduce several natural conditions on each player's payoff functions.

First is the condition of *Anonymity*, often employed in the literature, which requires that each player's payoff depends only on the number of players who choose each strategy.

Assumption Anonymity (AN). For any player $i \in N$, for any $S, T \subset N$ such that $i \in S \cap T$, and $|S| = |T|$, the equality $u^i(x, S) = u^i(x, T)$ holds for every alternative $x \in X$, where $|B|$ denotes the cardinality of the set *B*.

Condition *AN* allows us to use the notation

$$u^i(x, S) = h^i(x, |S|)$$

for every player $i \in N$, every subset *S* of *N* with $i \in S$ and every alternative $x \in X$.

The condition of *order-invariance* requires that players' preferences over alternatives are independent of the number of players choosing them as long as the number of players choosing these alternatives is the same. In the context of the example of a word processor choice in the department, mentioned in the introduction, order invariance implies that if an individual prefers *Word* over *Word Perfect* when she is the only word processor user, then she would still prefer *Word* over *Word Perfect* if both processes are used by the same number of departmental members. Formally.

Assumption Order-Invariance (OI). For any player $i \in N$, for any $S \subset N$ with $i \in S$ and any two alternatives $x, y \in X$, the inequality $u^i(x, S) \geq u^i(y, S)$ holds if and only if $u^i(x, \{i\}) \geq u^i(y, \{i\})$

Note that the *OI* is stronger than the no spillover condition. Thus, the set of games satisfying *OI* is the subset of the class of no spillover games. We now show that the conditions *PE, AN*, and *OI* do not guarantee the existence of a Nash equilibrium in the no spillover games. Indeed, consider the following example:

EXAMPLE 3.1. Let $N = \{1, 2, 3, 4, 5, 6\}$, and $X = \{a, b, c\}$. Let the players' payoff functions satisfy the following inequalities

$h^1(c, 4) > h^1(a, 3) > h^1(a, 2) > h^1(b, 3) > h^1(c, 3) > h^1(b, 2)$;
$h^2(b, 3) > h^2(a, 3) > h^2(b, 2) > h^2(b, 1) > h^2(a, 2) > h^2(a, 1)> h^2(c, k)$ for $k = 1, \dots , 6$;
$h^3(b, 2) > h^3(c, 4) > h^3(c, 3) > h^3(b, 1) > h^3(c, 2) > h^3(c, 1) > h^3(a, k)$ for $k = 1, \dots, 6$;
$h^4(a, 1) > h^4(x, k)$ for $x \neq a$; $k = 1, \dots, 6$;
$h^i (c, 1) > h^i (x , k)$ for $i = 5,6$; $x \neq c$; $k = 1, \dots , 6$.

Then the game G does not admit a Nash equilibrium.

Proof. Suppose, in negation that game G in this example possesses Nash equilibrium $(x^1, x^2, x^3, x^4, x^5, x^6)$. Then $x^4= a$, $x^5 = x^6 = c$. Since $x^3 \neq a$, it remains to consider two cases: $x^3 = b$ and $x^3 = c$.

Let $x^3 = b$. Since $x^2 \neq c$, there will be no four players choosing c, implying that player 1 should choose $x^1 = a$. Then the best response of player 2 is a. The strategy profile (a, a, b, a, c, c) is, however, not a Nash equilibrium as player 3 would rather choose c given the choices of the other five players.

Let $x^3 = c$. Then player 1 will choose $x^1 = c$. Then the best response of player 2 is b. The strategy profile (c, b, c, a, c, c) is, however, not a Nash equilibrium as player 3 would rather choose b given the choices of other five players. Thus, there is no Nash equilibrium in this game. ■

It is easy to see the constraints on the payoff functions in Example 3.1 do not violate *PE*, *AN*, *OI* and it is straightforward to define the payoff functions for all players so that all those conditions are satisfied. We would like to stress here that although each of the players 4, 5, and 6 has a strictly dominant strategy (alternative a for player 4 and alternative c for players 5 and 6), their presence is necessary to satisfy assumption *OI* which would be violated if these players are eliminated from the game. Indeed, let players 4, 5, and 6 be removed and introduce the payoff function $\tilde{h}^1(x,k)$ of player 1 in a way which takes into account their removal. Then we have $\tilde{h}^1(a,k) = \tilde{h}^1(a,k+1)$ and $\tilde{h}^1(c,k) = \tilde{h}^1(c,k+2)$ for all positive numbers k, yielding $\tilde{h}^1(c,2) > \tilde{h}^1(a,2)$ and $\tilde{h}^1(a,1) > \tilde{h}^1(c,1)$ so that *OI* is violated.

It is interesting to note that the preferences of each individual i in this example are single-peaked for any given group of individuals who choose the same strategy as i. That is, for every group of players S containing i, the function $u^i(\cdot, S)$ is single-peaked[5] over the set of alternatives X. Thus adding the single-peakedness condition (which is common in local public good economies where the

[5] Formally, the single-peakedness in our context is defined as follows: Let an ordering $\preceq$ on the set of alternatives X be given. The preferences of player $i \in N$ are *singled-peaked with respect to* $\preceq$ if for $S \subset N$ with $i \in S$, there exists an alternative x_i^S satisfying the following property: for any pair of alternatives $x, y \in X$ with either $x \preceq y \preceq x_i^S$ or $x \succeq y \succeq x_i^S$ it follows that $u^i(x, S) \leq u^i(y, S) \leq u^i(x_i^S, S)$. Then the preferences of every player $i \in N$ are *singled-peaked* if there exists an order ing $\preceq^*$ on X such that the preferences of every player $i \in N$ are singled-peaked with respect to $\preceq^*$

individuals' preferences are often assumed to be single-peaked with respect to the quantities of public goods produced in a given jurisdiction to assumptions *PE*, *AN*, and *OI* still does not guarantee the existence of a Nash equilibrium of a no spillover game.

4. Main Result

This example in the previous section shows that even under assumptions listed in the previous section, a Nash equilibrium still might fail to exist. The situation is quite different in the *negative externality* (*NE*) case, where payoff of any given player declines when a larger number of players choose that player's strategy.[6] Indeed, Milchtaich (1996), Konishi et al. (1997a), and Quint and Shubik (1994) show that in the models with *NE*, the anonymity assumption *AN* alone guarantees the existence of a Nash equilibrium in no spillover games. Konishi et al.(1997a) show that, moreover, under the same conditions, there exists even a strong Nash equilibrium.

Example 3.1 shows that if *NE* is replaced by *PE*, we cannot expect to have a Nash equilibrium even when assumption *OI* is imposed. Thus, we need to introduce an even stronger condition than *OI* on the regularity of a payoff function. *Order preservation* implies that if player i prefers alternative x to alternative y when players in N_x choose x and players in N_y choose y, then she would still prefer x over y if N_x and N_y are both expanded by an additional player j. Similarly, she would still prefer x over y if a common player k withdraws from both N_x and N_y.

Assumption Order Preservation (OP). For any $i, j \in N,$ for any $S, T \subset N$ such that $i \in S \cap T$ and $j \notin S \cup T,$ for any two alternatives $x, y \in X,$ $u^i(x, S) \geq u^i(y, T)$ *if and only if* $u^i(x, S \cup \{j\}) \geq u^i(y, T \cup \{j\})$.

Assumption *OP* is stronger than *OI*.[7] In fact, *OP*, together with *AN* implies that, for any integers l, m, r, if player i prefers alternative x chosen by l players over alternative y chosen by m players, then she would still prefer x chosen by $l + r$ players over y chosen by $m + r$ players.

Now we are in position to state our results on existence of a Nash equilibrium under *PE* when the set of alternatives consists of more than two elements. First, we consider the case where the set of alternatives X is finite.

[6] Formally, the condition of *negative externality* (NE) is defined as follows: For any two players $i, j \in N$, for any subset of players $S \subset N$ with $i \in S$ and $j \notin S$ and alternative $x \in X$ we have $u^i(x, S) \geq u^i(x, S \cup \{j\})$.

[7] Indeed, *OI* is equivalent to the following condition: For any $i, j \in N$, for any $S \subset N$ with $i \in S$ and $j \notin S$ for any two alternatives $x, y \in X$, $u^i(x, S) \geq u^i(y, S)$, if and only if $u^i(x, S \cup \{j\}) \geq u^i(y, S \cup \{j\})$. Thus, by setting $S = T$ in *OP*, it is easy to verify that *OP* is stronger than *OI*.

Proposition 4.1. *Suppose that X is finite and the payoff function of each player satisfies PE, AN, and OP. Then a no spillover game G admits a Nash equilibrium.*[8]

To prove this proposition, we make use of Lemma 4.2, the proof of which is presented in Section 6. Lemma 4.2 allows to eliminate, for each player i, the set of 'irrelevant' alternatives that would never be chosen by i in equilibrium, and to focus on the set of 'relevant' alternatives, X^i, that could be potential choices for i's equilibrium strategies. Moreover, Lemma 4.2 provides us with a quasi-linear utility representation theorem of each player i's payoff function over the 'relevant' set X^i. This utility representation plays the central role in the proof of Proposition 4.1.

Lemma 4.2. *Let the set of alternatives X be finite and assume that PE, AN, and OP hold. Then for every $i \in N$ there exists a nonempty set $X^i \subset X$ such that*

(i) *for any $x, y \in X^i$, $h^i(x, 1) \leq h^i(y, n)$,*
(ii) *for any $x \in X \setminus X^i$, there exists $y \in X^i$ such that $h^i(y, 1) \geq h^i(x, n)$.*

Moreover, there is a utility representation v^i: $X^i \rightarrow \Re$ such that one of the following two statements is true:

(iii) *for any $x \in X^i$, for any integer $1 \leq k \leq n$, $v^i(x) = h^i(x, k)$,*
(iv) *for any $x, y \in X^i$, for any integers k, m such that $1 \leq k$, $m \leq n$, $v^i(x) + k \geq v^i(y) + m$, if and only if $h^i(x, k) \geq h^i(y, m)$.*

Condition (i) states that for each player i no alternative in X^i chosen by i alone would be preferred over any other alternative in X^i when chosen by all players. Condition (ii) implies that for each alternative x, which is not in X^i, there exists an alternative y in X^i such that player i would weakly prefer y chosen alone over x when chosen by all other players. Conditions (iii) and (iv) provide the utility representation result.

Proof of Proposition 4.1. We shall prove this proposition assuming the validity of Lemma 4.2. Let $i \in N$. Let X^i be the set which is defined in Lemma 4.2. Let L and M be the sets of players whose payoff functions satisfy the conditions given by representation (iii) and (iv), respectively, of Lemma 4.2. Since the utility function v^i is defined over the set X^i, it follows that the sets L and M represent a partition of N, i.e., $L \cap M = \emptyset$, and $L \cup M = N$. Assign each player $i \in L$ to her best alternative (arg $\max_{x \in X}$, $v^i(x)$). Note that if these players are assigned to their best alternatives, they

[8] Note that we do not impose single-peakedness of individuals' preferences, which is frequently used in order to prove the existence of an equilibrium in local public good economies and models of multiparty electoral spatial competition.

would have no incentive to move to any other alternative regardless of all other players' choices. Let the resulting partition of players in L over X be $(L_x)_{x \in X}$, where $L_x \subset \{i \in L | \arg\max_{y \in X^i}, v^i(y) = x \text{ for all } x \in X\}$. (In the case where a player has more than one best alternative, we arbitrarily assign her to one of her best alternatives).

Now let us assign players in M to alternatives in X. We shall call $Q = (M_x)_{x \in X}$ a *legitimate* partition of players in M over X if for any $x \in X$ and any $i \in M$, $i \in M_x$ implies $x \in X^i$. That is, each player i is assigned only to one of her 'relevant' alternatives. Let Q be a collection of all legitimate partitions of M. Let $Q^* = (M_x^*)_{x \in X} \in Q$ be the legitimate partition of M that solves the following maximization problem:[9]

$$Q^* \in \arg\max_{Q \in Q} \sum_{x \in X} \left[\sum_{i \in M_x} v^i(x) + \sum_{k=1}^{|M_x|} (k + |L_x|) \right].$$

Since both the set of players N and the set of alternatives X are finite, Q^* is well defined. Let $\mathbf{x}^*$ be the strategy profile in which each alternative $x \in X$ is chosen only by the players who belong to the union of two sets, M_x^* and L_x, i.e., $P(\mathbf{x}^*) = ((M_x^* \cup L_x)_{x \in X})$. We shall show that the strategy profile $\mathbf{x}^*$ is a Nash equilibrium of game G. Otherwise, there exists a player j in M who would benefit from switching from one of her 'relevant' alternatives to another . (Note that no player $i \in M$ has an incentive to move to any alternative outside of X^i. That is, there is $j \in M$ and two alternatives a, $b \in X^j$ such that $j \in M^*$ and $v^j(a) + |M^*| + |L_a| < v^j(b) + |M_b^*| + |L_b| + 1$. Thus, we have the inequality

$$\begin{aligned}
&\sum_{x \in X} \left[\sum_{i \in M_x^*} v^i(x) + \sum_{k=1}^{|M_x^*|} (k + |L_x|) \right] \\
&\quad < \sum_{x \in X} \left[\sum_{i \in M_x^*} v^i(x) + \sum_{k=1}^{|M_x^*|} (k + |L_x|) \right] + (v^j(b) + |M_b^*| + |L_b| + 1) - (v^j(a) + |M_a^*| + |L_a|) \\
&\quad = \sum_{x \in X} \left[\sum_{i \in M'_x} v^i(x) + \sum_{k=1}^{|M'_x|} (k + |L_x|) \right]
\end{aligned}$$

where $M'_a = M_a^* \backslash \{j\}$, $M'_b = M_b^* \cup \{j\}$, and $M'_x = M_x^*$ for all $x \in X \backslash \{a, b\}$. Thus, the legitimate partition of M, $Q' = (M_x')_{x \in X} \in Q$ generates a higher value of the objective function than partition Q^*, a contradiction. Thus, $\mathbf{x}^*$ is a Nash equilibrium of the game G. ■

Although, in contrast to Rosenthal (1973), our game is not symmetric, the

[9] Note that the function in the bracket is not the sum of the payoffs of the players who choose strategy x. In fact, given our payoff representation in Lemma 4.2, the function provided here turns out to be a modification of the Monderer and Shapley (1996) exact potential function of the game G.

method of the proof of Proposition 4.1 is similar to the one used by Rosenthal who introduced the class of 'potential games' studied in Monderer and Shapley (1996). It is also important to note that our technique could be used to prove the existence of a Nash equilibrium in the games with a more general class of players' preferences under restrictions on players' preferences profiles (see Section 5).

We would also like to point out that the finiteness of set X, imposed in Proposition 4.1, is not essential. We can easily extend our result to the case with an infinite set of alternatives without requiring continuity of payoff functions.

Proposition 4.3. *Suppose that for each player $i \in N$ there exists an alternative $a_i \in X$, such that $h^i(a_i, 1) \geq h^i(x, 1)$ for any $x \in X$. Then, under PE, AN, and OP, the no spillover game G admits a Nash equilibrium.*

Proof. In order to show the existence of a Nash equilibrium of the game G with the infinite set of alternatives X we shall select a finite subset $\tilde{X}$ of the set of alternatives X such that the Nash equilibrium of the game $\tilde{G} = (N, \tilde{X}, U)$, the existence of which is guaranteed by Proposition 4.1, also constitutes a Nash equilibrium of the game $G = (N, X, U)$.

Indeed, for each player i let a_i be the i's top choice when chosen unilaterally, i.e., $a^i = \arg\max_{x \in X} h^i(x, 1)$. Let $\tilde{X} = \{(a_i)_{i \in N}\}$. By Proposition 4.1, the game $\tilde{G} = (N, \tilde{X}, U)$ admits a Nash equilibrium, denoted by $\mathbf{x} = (x^1, x^2, \ldots, x^n)$. Thus, *PE* implies that $U^i(x) \geq h^i(a_i, 1)$ for every player i. If x is not a Nash equilibrium of the game $G = (N, X, U)$, there exists a player i and an alternative $x \notin \tilde{X}$ such that $U^i(\mathbf{x}) \geq h^i(x, 1)$. Thus, $h^i(x, 1) > h^i(a_i, 1)$, a contradiction to the choice of a_i. ■

Example 3.1 above demonstrates that the assertion of Propositions 4.1 and 4.3 might not hold if *OP* is relaxed. We shall show that if *AN* is dropped, a Nash equilibrium might fail to exist even when the conditions *PE* and *OP* are satisfied:[10]

EXAMPLE 4.4. Let $N = \{1, 2, 3, 4, 5, 6\}$, and $X = \{a, b, c\}$. For each player $i = 1, 2, 3$ there exist functions $v^i: X \rightarrow \Re$ and $W^i: N \backslash \{i\} \rightarrow \Re$, which determ ines the 'value' of player $j \neq i$ for player i, such that the payoff function $u^i(x, S)$ is given by

$$u^i(x, S) = v^i(x) + \sum_{j \in S \backslash \{i\}} W^i(j).$$

The functions $v^i(\cdot)$ and $W^i(\cdot)$ assume the values

$v^1(a) = 6,$ $\quad$ $v^1(b) = 2\frac{1}{2},$ $\quad$ $v^1(c) = 0;$

[10] Note that in all examples of this section the players' preferences are single-peaked.

$W^1(2) = W^1(4) = 1,\qquad W^1(3) = W^1(5) = W^1(6) = 3;$

$v^2(\mathrm{a}) = 0,\qquad v^2(b) = 2,\qquad v^2(c) = -8;$

$W^2(1) = 3,\qquad W^2(j) = 1\qquad$ for $j = 3, 4, 5, 6;$

$v^3(a) = -8,\qquad v^3(b) = 1,\qquad v^3(c) = 0;$

$W^3(2) = 3,\qquad W^3(j) = 1\qquad$ for $j = 1, 4, 5, 6.$

The payoff functions of other three players satisfy the following inequalities:

$u^4(a, \{4\}) > u^4(x, N)\qquad$ for $x = b, c;$

$u^i(c, \{i\}) > u^i\,(x, N)\qquad$ for $i = 5, 6; x = a, b.$

Then the game G does not admit a Nash equilibrium.

Proof. The verification of conditions *PE* and *OP* is straightforward. It is also easy to check that the following inequalities are satisfied:

$$\begin{aligned}
u^1(c, \{1, 3, 5, 6\}) &> u^1(a, \{1, 2, 4\}) > u^1(a, \{1, 4\}) > u^1(b, \{1, 2, 3\})\\
&> u^1(c, \{1, 5, 6\}) > u^1(b, \{1, 3\}) > u^1(b, \{1, 2\})\ ;\\
u^2(b, \{1, 2, 3\}) &> u^2(b, \{1, 2\}) > u^2(a, \{1, 2, 4\}) > u^2(b, \{2, 3\})\\
&> u^2(b, \{2\}) > u^2(a, \{2, 4\}) > u^2(c, N)\ ;\\
u^3(b, \{1, 2, 3\}) &> u^3(b, \{2, 3\}) > u^3(c, \{1, 3, 5, 6\}) > u^3(c, \{3, 5, 6\})\\
&= u^3(b, \{1, 3\}) > u^3(b, \{3\}) > u^6\,(a, N).
\end{aligned}$$

It is easy to see that this example has the same structure as Example 3.1. Thus, one can use the arguments used there in order to show the nonexistence of a Nash equilibrium. ■

To demonstrate that *PE* cannot be dropped either, we construct the game with two players which satisfies *AN*, *OP* (and, trivially, single-peakedness) but does not admit a Nash equilibrium. In this game the preferences of the first player exhibit increasing returns to scale whereas the preferences of the second exhibit decreasing returns to scale (thus, violating *PE*).

EXAMPLE 4.5 – *Matching pennies*. Let $N = \{1, 2\}$, and $X = \{a, b\}$. Let players' payoff functions satisfy the following inequalities: $h^1(a, 2) = h^1(b, 2) > h^1(a, 1) = h^1(b, 1)$ and $h^2(a, 1) = h^2(b, 1) > h^2(a, 2) = h^2(b, 2)$. Then the game G does not admit a Nash equilibrium.

Proof. The verification of conditions *AN* and *OP* is straightforward. If both players choose the same alternative, player 2 would be better off by switching to a different alternative as she would rather stay alone. If the players choose different alternatives, player 1 would be better off by switching to the alternative chosen by player 2. Thus, this game does not admit a Nash equilibrium. ■

The next question which arises naturally is whether, similarly to Greenberg and Weber (1986, 1993), the existence of strong Nash equilibrium is guaranteed under the same assumptions which yield the existence of a Nash equilibrium. Weber and Zamir (1985) show that in the second-best local public good economy studied in Guesnerie and Oddou (1981) a strong Nash equilibrium may fail to exist.[11] The Weber and Zamir example, however, violates *OP*. The game in Example 4.6 below which does not admit a strong Nash equilibrium,[12] demonstrates that even *PE*, *AN*, *OP*, and single-peakedness are not, in general, sufficient to yield the existence of a group structure which is stable under group deviations.

Example 4.6. Let $N = \{1, 2, 3, 4, 5, 6, 7\}$ and $X = \{a, b, c\}$. For every alternative $x \in X$ and every integer k, $1 \le k \le 6$, the value of player *i*'s payoff function is represented by $h^i(x, k) = v^i(x) + k$, where

$$\begin{array}{lll}
v^1(a) = -1\,\tfrac{1}{2}, & v^1(b) = 0, & v^1(c) = -3\,\tfrac{7}{10}; \\
v^2(a) = -3\,\tfrac{7}{10}, & v^2(b) = \tfrac{1}{2}, & v^2(c) = -2; \\
v^3(a) = -2, & v^3(b) = -1\,\tfrac{7}{10}, & v^3(c) = -1\,\tfrac{1}{2}; \\
v^i(a) = 0, & v^i(b) = v^i(c) = -8 & \text{for } i = 4, 5; \\
v^i(a) = v^i(b) = -8, & v^i(c) = 0 & \text{for } i = 6, 7.
\end{array}$$

Then the game G does not admit a strong Nash equilibrium.

Proof. It is trivial to check that the payoff functions are single-peaked and satisfy *PE*, *AN*, *OP*.

Note that at any Nash equilibrium players 4 and 5 would choose alternative a, whereas players 6 and 7 would choose alternative b. Consider the reduced game $\tilde{G}$ obtained by 'fixing' the choices of players 4 and 5 at a, players 6 and 7 at c and accordingly adjusting the payoff functions of players 1, 2, and 3 in game G. Formally, the game $\tilde{G}$ is given by the set of players $\tilde{N} = \{1, 2, 3\}$, the set of alternatives $X = \{a, b, c\}$ and the players' payoff functions determined by the parameters:

[11] It is easy to verify that the Weber and Zamir example admits a Nash equilibrium. Konishi et al. (1995) demonstrate, however, that the existence of a pure strategy Nash equilibrium is not, in general, guaranteed in the Guesnerie and Oddou economy with more than three agents.

[12] Konishi et al. (1997) show that if *PE* is satisfied, then the set of coalition-proof and strong Nash equilibria of the no spillover game coincide. This result implies that the game in Example 4.6 does not admit a coalition-proof Nash equilibrium either.

$$\begin{array}{lll} \hat{v}^i(a) = \frac{1}{2}, & \tilde{v}^1(b) = 0, & \tilde{v}^1(c) = -1\frac{7}{10}; \\ \tilde{v}^2(a) = -1\frac{7}{10}, & \tilde{v}^2(b) = \frac{1}{2}, & \tilde{v}^2(c) = 0; \\ \tilde{v}^3(a) = 0, & \tilde{v}^3(b) = -1\frac{7}{10}, & v^3(c) = \frac{1}{2}. \end{array}$$

The payoff functions of the game $\tilde{G}$ satisfy *PE*, *AN*, and *OP*, but violate single-peakedness. Thus, four additional players were added to construct the game G which would satisfy the single-peakedness assumption as well. It is easy to see that a strategy profile (x^1, x^2, x^3, a, a, b, b) constitutes a strong Nash equilibrium of the game G if and only if the triple (x^1, x^2, x^3) constitutes a strong Nash equilibrium of the game $\tilde{G}$. It remains, therefore, to show that $\tilde{G}$ does not admit a strong Nash equilibrium. Indeed, the game $\tilde{G}$ admits three Nash equilibria, $\mathbf{x} = (a, b, a)$, $\mathbf{y} = (a, c, c)$ and $\mathbf{z} = (b, b, c)$. However,

at $\mathbf{x}$, players 2 and 3 would be better off by jointly switching to c,
at $\mathbf{y}$, players 1 and 2 would be better off by jointly switching to b,
at $\mathbf{z}$, players 1 and 3 would be better off by jointly switching to a.

Thus, the game $\tilde{G}$ does not admit a strong Nash equilibrium. ■

5. Extension

The primary focus of this paper is to identify a domain of preferences such that any game where the preferences of each player drawn from this domain possesses a Nash equilibrium. However, we can use our technique to prove the existence of a Nash equilibrium in a wider class of games which allow for restrictions on preferences profiles rather than on domain of individual preferences.[13] Indeed, the method of the proof applied in Proposition 4.1 can be used to obtain the existence of a Nash equilibrium in a class of games where the payoff function of each player i is given by

$$u^i(x,S) = v^i(x) + \phi(x, |S|),$$

where S is the set of players (including i) who choose the strategy x and the function $\phi(\cdot)$ represents the value of the externality effect which is common for all players.[14] Consider, for example, an environment with network externalities and consider the set S of all those customers who choose alternative x. Then for every

[13] We thank the associate editor and the referee for the suggestion to add the discussion on possible extensions of our main result.

[14] The only modification required to accommodate this class of preferences is to replace the number k in the proof of Proposition 4.1 by the value of the common function $\phi(x, k)$. It is useful to observe that the set L_x in the proof would be empty for every strategy $x \in X$.

player $i \in S$, the function $\phi(x, |S|)$ represents the (common) gain of i generated from the fact that the alternative x has been chosen by all members of S.[15] It is important to point out that while the externality effect could depend both on the choice of alternative (x) and the size of the group that chooses it, ($|S|$), the effect *must* be identical for all members of S. This requirement is, obviously, much stronger than the assumption *AN*. However, since the function $\phi(x, \cdot)$ does not have to be linear and may, for example, exhibit decreasing returns to scale with respect to the number of individuals who choose x, the assumption *OP* does not necessarily hold.

We complete this section by showing that 'commonality' of the externality effect is crucial to obtain the existence of a Nash equilibrium. The next example shows that if the externality effect is not common for all players, then even in the case where the payoff functions are of a special separable functional form, satisfying the no-spillover condition and assumptions *PE* and *AN*, a Nash equilibrium may fail to exist.

EXAMPLE 5.1. Let $N = \{1, 2, 3, 4, 5, 6\}$ and $X = \{a, b, c\}$. Let external effect be independent of the choice of alternative and the preferences of each player $i \in N$ be given by

$$u^i(x, S) = v^i(x) + \phi^i(|S|),$$

where

$v^1(a) = 1\frac{2}{3}$,	$v^1(b) = \frac{1}{2}$,	$v^1(c) = 0$;	
$\phi^1(1) = \phi^1(2) = 0$,	$\phi^1(3) = 1$,	$\phi^1(4) = \phi^1(5) = \phi^1(6) = 3$;	
$v^2(a) = \frac{1}{2}$,	$v^2(b) = = 1\frac{2}{3}$,	$v^2(c) = -3$;	
$\phi^2(1) = 0$,	$\phi^2(2) = 1$,	$\phi^2(3) = \phi^2(4) = \phi^2(5) = \phi^2(6) = 3$;	
$v^3(a) = -4$;	$v^3(b) = 2\frac{1}{3}$,	$v^3(c) = 0$;	
$\phi^3(1) = 0$,	$\phi^3(2) = 2$,	$\phi^3(3) = 3$,	$\phi^3(4) = \phi^3(5) = \phi^3(6) = 3\frac{1}{2}$,
$v^4(a) = 1$,	$v^4(b) = v^4(c) = 0$;		
$\phi^3(k) = 0$,	for all $k = 1, \ldots, 6$;		
$v^i(a) = v^i(b) = 0$,	$v^i(c) = 1$;		
$\phi^i(k) = 0$,	for $i = 5, 6$; for all $k = 1, \ldots, 6$.		

Then this game does not admit a Nash equilibrium.

The structure of payoff functions is exactly the same as in Example 3.1 and it

[15] See Konishi et al. (1995) for a discussion on interpretation of this preferences' specification in the context of economies with club or local public goods.

is easy to see, therefore, that the assumptions *PE*, *AN*, and even *OI* hold; all preferences are single-peaked, whereas the set of Nash equilibria is empty. Note that in this example the externality effect $\phi^i(|S|)$ is even independent of an alternative chosen by players in S. This highlights the importance of common externality effect for the existence of a Nash equilibrium.

Appendix

Proof of Lemma 4.2. For each player $i \in N$ denote by

$$X_0^i = \{x \in X \mid h^i(x,n) \geq h^i(y,n) \text{ for all } y \in X\}$$

the set of best alternatives for $i \in N$. Since X is finite, every $X_0^i \neq \varnothing$. For each i denote

$$X^i = \{x \in X \mid h^i(x,n) > h^i(y,1) \text{ for all } y \in X_0^i\} \cup X_0^i$$

By the construction, the sets X^i satisfy the first two assertions of Lemma 4.2.

To prove the last two assertions of Lemma 4.2, we shall make use of the Konishi and Fishburn (1996) utility representation theorem:[16]

Proposition A.1. *Suppose that OP is satisfied and the following two conditions hold:*

(1) for any $x \in X^i$, for any integer $1 \leq k \leq n - 1$, we have $h^i(x, k) < h^i(x, k + 1)$,
(2) for any $x, y \in X^i$, there exists an integer $1 \leq n_{yx} \leq n$ such that $h^i(y, n_{yx}) > h^i(x,1)$

Then, for every $i \in N$, there exists a function v^i: $X^i \rightarrow \Re$, such that for any pair of integers $1 \leq m, k \leq n$, and any pair of alternatives $x, y \in X^i$, the inequality $v^i(x) + m \geq v^i(y) + k$ holds if and only if $h^i(x, m) \geq h^i(y, k)$.

To apply this result to Lemma 4.2, we need the following.

Lemma A.2. *Under PE, AN, and OP, one of the following statements is true:*

(α) for any $x \in X^i$, and any integer k, $1 \leq k \leq n - 1$, $h^i(x, k) < h^i(x, k + 1)$,
(β) for any $x \in X^i$, for any two integers k, m with $1 \leq k, m \leq n$, $h^i(x, k) = h^i(x, m)$.

Proof. Take any $i \in N$. Let first X^i consist of a single element, x. By *PE* and *OP*, if $h^i(x, 1) < h^i(x, n)$ then (α) holds, and if $h^i(x, 1) = h^i(x, n)$ then (β) holds.

Suppose now that $|X^i| \geq 2$. If for any $x, y \in X^i$, $h^i(x, 1) = h^i(y, 1)$, then, by *OP*,

[16] We provide here a slightly modified version of their result.

the equality $h^i(x, k) = h^i(y, k)$ holds for any integer k, $1 \leq k \leq n - 1$. Hence, all alternatives in X^i are essentially equivalent, and the argument is the same as in the previous case. Suppose, therefore, that there exist two alternatives $x, y \in X^i$ such that $h^i(x, 1) = h^i(y, 1)$. We shall show that only (α) can occur. First, consider alternative y. By (i) in Lemma 4.2, $h^i(x, 1) \leq h^i(y, n)$ Thus, $h^i(y, 1) < h^i(y, n)$. Condition *OP* implies that $h^i(y, k) < h^i(y, k + 1)$ for any integer $1 \leq k \leq n - 1$.

Consider now alternative x. Again, by *OP*, there exists an integer n_{yx} such that $2 \leq n_{yx} \leq n$, $h^i(y, n_{yx}) \geq h^i(x, 1)$, and $h^i(x, 1) > h^i(y, n_{yx} - 1)$. By *OP*, $h^i(x, 2) > h^i(y, n_{yx})$. Thus, $h^i(x, 2) > h^i(x, 1)$. By *OP* again, we conclude that $h^i(x, k) < h^i(x, k + 1)$ for any integer $1 \leq k \leq n - 1$.

Finally, let $z \in X^i \setminus \{x, y\}$. There are three cases: $h^i(z, 1) = h^i(y, 1)$, $h^i(z, 1) > h^i(y, 1)$, and $h^i(z, 1) < h^i(y, 1)$. The case where $h^i(z, 1) = h^i(y, 1)$ is trivial. The case $h^i(z, 1) > h^i(y, 1)$ can be treated in the same manner as the proof of the inequality $h^i(x, k) < h^i(x, k + 1)$ for any integer $1 \leq k \leq n - 1$. In the case where $h^i(z, 1) < h^i(y, 1)$ we have, by assertion (i) in Lemma 4.2, $h^i(y, 1) \leq h^i(z, n)$. Thus, $h^i(z, 1) < h^i(z, n)$ which, by *OP*, yields $h^i(z, k) < h^i(z, k + 1)$ for any integer k, $1 \leq k \leq n - 1$. ■

To complete the proof of Lemma 4.2, it remains to observe that, by Proposition A.1, (α) and (β) in Lemma A.2 yield assertions (iii) and (iv) of Lemma 4.2, respectively. ■

References[17]

Arthur, W.B. (1989), 'Competing technologies, increasing returns, and lock-in by historical events', *Economic Journal* **99**, 116–131.

Aumann, R.J. (1959), 'Acceptable points in general cooperative n-person games', in *Contributions to the Theory of Games*, Vol. IV, Princeton, NJ: Princeton Univ. Press.

Bernheim, D., B. Peleg and M. Whinston (1987), 'Coalition-proof Nash equilibria: I concepts', *Journal of Economic Theory* **42**, 1–12.

Demange, G. (1994), 'Intermediate preferences and stable coalition structures', *Journal of Mathematical Economics* **23**, 45–58.

Farrell, J. and G. Saloner, (1985), 'Standardization, compatibility, and innovation', *RAND Journal of Economics* **16**, 70–83.

Farrell, J. and G. Saloner (1988), 'Coordination through committees and markets', *RAND Journal of Economics* **19**, 235–252.

Greenberg, J. and S. Weber (1986), 'Strong tiebout equilibrium under restricted preferences domain', *Journal of Economic Theory* **38**, 101–117.

Greenberg, J. and S. Weber (1993), 'Stable coalition structures with unidimensional set of alternatives', *Journal of Economic Theory* **60**, 693–703.

Guesnerie, R. and C. Oddou (1981), 'Second best taxation as a game', *Journal of Economic Theory* **25**, 67–91.

[17] The references Konishi et al. (2007a) and Konishi et al. (2007b) were originally cited as forthcoming have been updated.

Katz, M.L. and C. Shapiro (1985), ‘Network externalities, competition, and compatibility’, *American Economic Review* **75**(3), 424–440.

Konishi, H. and P. Fishburn (1996), ‘Quasi-linear utility in a discrete choice model’, *Economics Letters* **51**, 197–200.

Konishi, H., M. Le Breton and S. Weber (1995), ‘Equilibrium in a finite local public goods economy’, Southern Methodist University Working Paper.

Konishi, H., M. Le Breton and S. Weber (1997a), ‘Equilibrium in a model with partial rivalry’, *Journal of Economic Theory* **72**(1), 225–237.

Konishi, H., M. Le Breton and S. Weber (1997b), ‘Equivalence of strong and coalition-proof Nash equilibria in games without spillovers’, *Economic Theory* **9**(1), 97–113.

Liebowitz, S.J. and S.E. Margolis (1994), ‘Network externality: An uncommon tragedy’, *Journal of Economic Perspectives* **8**(2), 137–151.

Milchtaich, I. (1996), ‘Congestion games’, *Games and Economic Behavior* **13**, 111–124.

Monderer, D. and L. Shapley (1996), ‘Potential games’, *Games and Economic Behavior* **13**, 124–143.

Quint, T. and M. Shubik (1994), ‘A model of migration’, mimeo, Yale University.

Rosenthal, R.W. (1973), ‘A class of games possessing a pure-strategy Nash equilibrium’, *International Journal of Game Theory* **2**, 65–67.

Tirole, J. (1988), *Industrial Organization*, Cambridge, MA: The MIT Press.

Weber, S. and S. Zamir (1985), ‘Proportional taxation: Nonexistence of stable structures in an economy with a public good’, *Journal of Economic Theory* **35**, 178–185.

A Theory of Full International Cooperation

Scott Barrett

Paper presented at the CTN Workshop 1998. Reprinted from *Journal of Theoretical Politics* **11**(4), October 1999, 519–541 with permission from SAGE Journals.

This paper develops a coherent theory of international cooperation relying on the twin assumptions of individual and collective rationality. Using a linear version of the N-player prisoner's dilemma game, I provide a formal proof of Olson's conjecture that only a 'small' number of countries can sustain full cooperation by means of a self-enforcing agreement. Moreover, I find that this number is not fixed but depends on the nature of the cooperation problem; for some problems, three countries will be 'too many', while for others even 200 countries will be a 'small' number. In addition, I find that the international system is only able to sustain global cooperation – that is, cooperation involving 200 or so countries – by a self-enforcing treaty when the gains to cooperation are 'small'. Finally, I find that the ability of the international system to sustain cooperation does not hinge on whether the compliance norm of customary international law has been internalized by states or whether compliance must instead be enforced by the use of treaty-based sanctions. The constraint on international cooperation is free-rider deterrence, not compliance enforcement.

1. Introduction

The theory of international cooperation developed in this paper assumes that cooperative arrangements between countries must be both individually and collectively rational: individually rational because the choice of whether to be a party to a treaty is voluntary; collectively rational because diplomats meet face to

I am grateful to Geir Asheim, Olivier Compte, Jeroen Hinloopen, and Marco Mariotti for helpful comments at conference and workshop presentations and to three anonymous referees and the editor of this journal for helping an economist learn how to write a political science paper

face and so can exploit fully the potential joint gains from cooperation in a treaty. Individual rationality is a standard assumption in the literature. Collective rationality is a more novel assumption, but it is compelling nonetheless. In this paper I show that the combination of these assumptions has profound implications for the theory of international cooperation.

Two pillars of the received theory are (1) that cooperation can be sustained as an equilibrium of a noncooperative repeated game by strategies of reciprocity (Axelrod, 1984; Axelrod and Keohane, 1985; Keohane, 1986); and (2) that cooperation can only be supported by a 'small' number of countries (Olson, 1965; Keohane, 1986). These features of the theory should be compatible but it is not obvious that they are. Indeed, the 'folk theorems' invoked to explain (1) clash with (2); they show that, for small enough discount rates, cooperation can be sustained as an equilibrium for any number of players. Olson supports the second pillar of the theory by a convincing, intuitive argument that appeals to the principle of reciprocity, but he does not offer a formal proof of the claim and nor, to my knowledge, has anyone else. So the two pillars remain unreconciled. However, the folk theorems rely only on the assumption of individual rationality; they do not require that agreements also be collectively rational. I show in this paper that the combination of these assumptions makes features (1) and (2) of the received theory compatible.

In particular, I provide a formal proof of Olson's (1965) conjecture that full cooperation can be sustained by means of a self-enforcing agreement only if the number of players is 'small'.[1] More than that, I show that whether any given number of countries is 'small' depends on the problem at hand. This means that full cooperation can sometimes be sustained by a great many countries and sometimes not even by a few. In showing this, I solve a puzzle in the literature: why some treaties can be sustained by nearly all the countries in the world when others cannot even be sustained by a handful of countries (see Keohane and Ostrom, 1994; Snidal, 1994; Young and Osherenko, 1993). Finally, I show what this means for world welfare. I find that there is an inverse relationship between the maximum number of countries that can sustain full cooperation by means of a self-enforcing agreement and the aggregate gains to cooperation. The international system, hampered as it is by the principle of sovereignty, can only sustain full cooperation among *all* the world's 200 or so countries when the total gains to cooperation are 'small' – that is, when a global agreement is not really needed. I demonstrate these points by analyzing a linear version of the symmetric prisoner's dilemma game, which captures the essentials of the cooperation problem and yet requires amazingly little mathematics. However, I emphasize that the basic insights of the paper can be shown to hold more generally.

[1] Of course, one can always limit cooperation in a repeated game by assuming that discount rates are high enough. I show that cooperation is limited, even for arbitrarily small discount rates.

What difference does the assumption of collective rationality make to the theory? Full cooperation can only be sustained by an international treaty if no country can gain by not being a party to it, and no party can gain by not implementing it. That is, free-riding must be deterred and compliance must be enforced. An agreement must therefore specify a strategy – a plan detailing what the parties should do – and this strategy, if obeyed, must succeed in deterring free-riding and enforcing compliance. Moreover, it must be in the interests of the parties actually to behave as the strategy demands. That is, the threat to reciprocate, to harm a country that has deviated from the strategy, must be credible. Essentially, the assumptions of individual and collective rationality define what we mean by a 'credible' strategy.

Individual rationality implies that, if every other country plays the equilibrium strategy, each can do no better than to play this strategy; and that, if a country did deviate from this strategy, then this country would want to revert to the equilibrium strategy and so would each of the others want to impose the punishment prescribed by the strategy, given that all other countries obeyed the strategy. That is, when push comes to shove, free-riding and noncompliance are punished; and it is precisely because it is known that this behavior will be punished that no country deviates in equilibrium.

Collective rationality, as the term is used in this paper, implies that an equilibrium agreement cannot be vulnerable to renegotiation. This means, first, that there cannot exist an alternative, feasible agreement that all countries prefer to the equilibrium agreement; and, second, that should a country deviate from the equilibrium, not only would this deviant want to revert to the equilibrium strategy, and not only would every other country behave in the manner prescribed by this strategy, given that all others did so, but all of the countries called upon to punish the defection would actually want to carry out the punishment and would not be tempted to renegotiate the agreement – to choose an alternative, feasible punishment or overlook the defection and not punish the defector at all. The agreements I consider are thus efficient in the sense that they sustain full cooperation by the threat of imposing efficient punishments.

Agreements seeking to sustain full cooperation are especially vulnerable to free-riding. This is because the greater is the extent of cooperation, the greater are the incentives to deviate – a point made by Downs et al. (1996). Hence, for a given number of countries, N, an agreement seeking to sustain full cooperation must impose a larger punishment to deter free-riding than an agreement seeking to sustain less cooperation. This by itself will make cooperation harder to sustain, since any punishment that is actually imposed harms the countries called upon to enforce the agreement as well as those on the receiving end. However, we know from the folk theorems of repeated games that for small enough discount rates even full cooperation can be sustained by a self-enforcing agreement.[2] I show in this paper that full cooperation cannot always be sustained by a self-enforcing agreement even when discount rates are arbitrarily small. Hence, it is the requirement that the punishments be renegotiation-proof; that is the crucial refinement and it is this that gives this paper its distinctive results.

In an infinitely repeated game, strategies capable of deterring a unilateral defection are credible (assuming that countries are sufficiently patient), if by 'credible' we mean that the strategies are individually rational. This is what the folk theorems tell us. But such strategies will not be credible (even for arbitrarily small discount rates) if by 'credible' we mean that they are collectively rational, provided N is large enough. The reason is that, the larger is N, the greater will be the harm suffered by the $(N-1)$ 'other' countries when they impose the punishment needed to deter a unilateral deviation. If N is large enough, it will not be in the collective interests of these countries actually to impose this punishment, should a deviation occur. An agreement which asks its signatories to play this 'incredible' strategy would be vulnerable to renegotiation; it would therefore not be self-enforcing.

As just indicated, my analysis is cast in a repeated game setting, and yet Chayes and Chayes (1995) have recently challenged the applicability of the theory of repeated games to problems of international cooperation. They claim that cooperation is sustained by the international compliance norm and not, as suggested by the theory of repeated games, treaty-based sanctions. The authority to impose sanctions, they note, 'is rarely granted by treaty, rarely used when granted, and likely to be ineffective when used' (Chayes and Chayes, 1995: 32–3). Downs et al. (1996; hereafter DRB) disagree that treaty-based sanctions are not needed. They argue that

> both the high rate of compliance and relative absence of enforcement threats are due not so much to the irrelevance of enforcement as to the fact that states are avoiding deep cooperation – and the benefits it holds whenever a prisoner's dilemma situation exists – because they are unwilling or unable to pay the costs of enforcement.
> (DRB, 1996: 387)

It is hard to take sides in this debate, because the Chayes's consider the compliance problem in isolation from free-riding, while DRB conflate these two problems.[3] Compliance and free-riding are different problems. But they are

[2] This may not be obvious from reading Downs et al. (1996), but note that in their example they assume a discount rate of 5 percent. As the discount rate is lowered, the weight of future punishments increases against the immediate gain to a defection. If the discount rate is low enough the folk theorems tell us that even full cooperation can be sustained as an equilibrium to a noncooperative game. With higher discount rates, cooperation will thus be harder to sustain than shown in this paper.

[3] Specifically, though DRB's paper is concerned only with compliance, their analysis of the incentives to defect can be interpreted as applying either to noncompliance or to non-participation. A defection is just a defection in their analysis, just as it is in my own repeated game model.

related problems and should be analyzed jointly. Doing so, however, poses an analytical problem: the theory of repeated games does not distinguish between 'defection' as a failure to *comply* with an agreement and 'defection' as a failure to *participate* in an agreement. The distinction is important, however, because while countries might be compelled, by means of the compliance norm of international law, to comply with the agreements they sign up to, there does not exist an international norm that requires that states *be* signatories to a cooperative agreement. Indeed, the essence of sovereignty is that states are free to participate in treaties or not as they please.

In the second half of this paper I recast the problem of international cooperation as a stage game in which signatories are assumed to choose their actions jointly so as to maximize their collective payoff (as required by collective rationality), in which nonsignatories are assumed to choose their actions independently so as to maximize their individual payoffs (as required by individual rationality), and in which all countries are free to be signatories or nonsignatories (as also required by individual rationality). As noted earlier, the Chayes's and DRB agree that countries comply with the agreements they sign up to; what they disagree on is whether this means that treaty-based sanctions are not needed and whether anything like deep cooperation can be sustained by the international system. I therefore adopt the tactic of assuming that all countries have internalized the compliance norm of customary international law in order to see whether this assumption matters.[4] I show that DRB are right that the international system may fail miserably at sustaining deep cooperation, even assuming that the Chayes's are right that the norms of international behavior suffice to ensure that countries comply fully with their international obligations. Like the earlier result, I also find that only a 'small' number of countries can sustain the full cooperative outcome, and that there is an inverse relationship between the maximum number of countries that can sustain full cooperation and the total gains to cooperation.

Because of their different formulations, the repeated and stage game models sustain cooperation by means of different strategies. To sustain full cooperation as an equilibrium of a repeated prisoner's dilemma, collective rationality requires that, if a party to an agreement plays Defect, the other parties can do no better collectively than to respond by playing Defect; and that, if this defector subsequently plays Cooperate in a punishment phase, to make amends for its earlier transgression, all the other parties to the agreement still can do no better collectively than to continue to play Defect – that is, to punish the original

[4] To assume that states have internalized the compliance norm is to assume that states will comply with an agreement they have signed up to, whether or not is in their interests to do so. This should be interpreted only as shorthand for the assumption that the compliance norm is sustained outside of the model under consideration. Kandori (1992) shows how norms can be sustained by community enforcement.

defection (it is this that makes the agreement 'renegotiation-proof'). To sustain full cooperation as an equilibrium of the stage game, collective rationality requires only that the first of these conditions be obeyed (the second cannot figure in the stage game model, because this game is essentially 'one-shot' and so there cannot exist a 'punishment phase'): that, if one country plays Defect, all the other countries can do no better collectively than to play Defect. Though different in the details, both strategies have the same basic requirement: that the countries responsible for punishing a unilateral defection must not be able to do better, either individually or collectively, by not carrying out the punishment specified in the treaty. Put differently, both approaches require that cooperation be enforced by credible punishment strategies.

Moreover, for a certain and important class of cooperation problem – one where the cost of participating in a treaty is independent of the number of countries that participate – I show that these conditions are identical. In other words, the compliance norm does not buy any additional cooperation.[5] The reason is intuitive. Any punishment to deter noncompliance must 'fit the crime.' So the larger the potential compliance failure is, the larger must be the threatened punishment if non-compliance is to be deterred. The greatest harm that any one signatory can inflict on the others is to do what it would do if it withdrew from the treaty entirely. So if a treaty can credibly threaten to impose a punishment that deters signatories from withdrawing unilaterally, it can easily threaten to impose a punishment that deters signatories from failing to comply with the agreement unilaterally. Once free-riding has been deterred, compliance enforcement comes free of charge.

This result needs to be modified slightly if the cost to each country of playing Cooperate is decreasing in the number of countries that play Cooperate – if there are increasing returns to cooperation. For, in comparison with the case discussed earlier, if any country plays Defect, the payoff to the others of playing Defect increases (punishing a defection becomes more attractive), whereas if a country plays Cooperate in a punishment phase, the payoff to the others of continuing to play Defect decreases (punishing a defector becomes less attractive). Increasing returns thus makes cooperation a little easier to sustain in the stage game model than in the repeated game model. But the reason for this is not that the

[5] This should not come as a surprise. In the model presented here, noncompliance implies that a signatory will play Defect when the agreement requires that it play Cooperate. So a signatory that fails to comply with the agreement will be indistinguishable from a country that free-rides on the agreement. However, it can be shown that if action sets are continuous and noncompliance can therefore entail only a slight deviation – a slight increase in pollution relative to the level prescribed by the agreement, for example – then the compliance norm will still deliver no additional cooperation. The reason is that if the agreement can deter a unilateral withdrawal, it can easily deter a lesser deviation. To deter a lesser deviation requires a smaller punishment, and smaller punishments harm the countries that are called upon to carry them out less than larger punishments. If a larger punishment is credible, therefore, so will be a smaller punishment.

assumption of full compliance buys any additional cooperation. The reason is that the stage game lacks a temporal dimension and so cannot specify explicitly an appropriate strategy of reciprocity.

The analysis developed in the paper is abstract. Many important features of real world cooperation problems like climate change mitigation and ozone layer protection do not figure in the model – to take two obvious examples, I assume that countries are symmetric and do not interact in other spheres so that issue linkage and reputation play no role here.[6] Moreover, the focus of my analysis is narrow. My interest is in determining the conditions that must hold for full cooperation to be sustained by the anarchic international system. I have little to say in this paper about whether something short of full cooperation can be sustained. But for all of these limitations, the theory is relevant to the real world, as the following example illustrates.

The Montreal Protocol sustains something very close to full cooperation. Nearly every country is a party to this agreement, and in implementing it the most harmful ozone-depleting substances are being phased out around the world. At a recent conference of the parties to the Montreal Protocol, delegates suggested (not for the first time) that this agreement should serve as a model for the climate change negotiations, which were soon to be convened in Kyoto. The analysis developed in this paper is useful for knowing whether their ambition could be met – whether the success at Montreal could be replicated in Kyoto. The theory tells us that it could be, but only if the underlying payoffs are favorable to international cooperation. Of course, these payoffs are givens, and so it may not be possible for the Kyoto negotiatiors to match the success of the Montreal Protocol.[7] To sustain full cooperation requires more than negotiation acumen, more than leadership, more than an active epistemic community, more even than an assurance that countries will obey the compliance norm. It depends also on whether the payoffs are of a magnitude that make the threat to punish deviations from full cooperation credible. This is the central message of this paper.

Before proceeding to the substance of the paper, I should perhaps comment on why I specialize by analyzing cooperation as an *international* problem. Certainly, the theory does have relevance to other problems. But the rules of the game of cooperation vary in different situations, and one must take care before extrapolating.[8] Where cooperation among firms is legal, it can be codified in a

[6] I discuss the implication of symmetry and issue linkage in the concluding section of the paper.

[7] As it happens, the agreement negotiated in Kyoto bears a number of similarities to the Montreal Protocol. Crucially, however, the Kyoto Protocol does not contain a free-rider deterrence mechanism, and – consistent with the insights of this paper – nor does it contain a non-compliance deterrence mechanism. The Montreal Protocol is different. It deters free- riding by means of trade sanctions between parties and nonparties. Trade sanctions have also been used to punish noncompliance with this agreement. See the concluding section of this paper.

[8] See the Special Issue of the *Journal of Theoretical Politics* 6(4), 1994.

contract, which can then be enforced by the courts having jurisdiction over the parties. Cooperative arrangements arrived at in this setting need not be self-enforcing. Where cooperation among firms is illegal, it may no longer be possible for firms to negotiate openly, and in this context the notion of collective rationality is less compelling. Finally, local, self-organized collective action problems of the type analyzed by Ostrom (1990) take place in settings where there is, at the very least, a potential for central intervention.[9] Context matters to the analysis of cooperation, and though the theory developed here will have implications for different settings, I apply it in this paper only to inter-state relations (and indeed only to a subset of these).

2. Individual Rationality in the One-Shot, *N*-Player Prisoners' Dilemma

The underlying game is assumed to be an N-player prisoners' dilemma, where $N \geq 2$, where countries must choose between playing Cooperate and Defect, and where the payoffs to each of the symmetric players of making these choices (Π_D and Π_C, respectively) are linear functions of the total number of countries that play Cooperate, z:

$$\Pi_D(z) = bz, \qquad \Pi_C(z) = -c + dz \tag{1}$$

In (1), b, c and d are parameters, and the payoffs have been normalized such that $\Pi_D(0) = 0$. This linear formulation is obviously special, but it will allow us to obtain very strong results using very little mathematics.

The prisoner's dilemma has three important features, and the parameters in (1) must be restricted to ensure that these are satisfied by the model.

The first feature of the prisoner's dilemma is that play Defect is a dominant strategy in the one-shot game. This means that every player must get a higher payoff when playing Defect than when playing Cooperate, irrespective of the number of other countries that play Defect (Cooperate). Formally, I limit my attention to problems that satisfy:

$$bz > -c + d(z+1) \qquad \text{for all } z,\ 0 \leq z \leq N-1 \tag{2}$$

The second feature of the prisoner's dilemma is that country i's payoff is increasing in the number of other countries that play Cooperate, irrespective of whether i plays Defect or Cooperate. This implies b, $d > 0$. Furthermore, upon setting $z = 0$ we see that (2) requires $0 > -c + d$, and so, given that $d > 0$, we must have $c > d$.

[9] For example, Ostrom (1990) begins her study by discussing the inshore fishery at Alanya, where the cooperative which developed rules for managing the community resource had previously been given jurisdiction over such matters by national legislation.

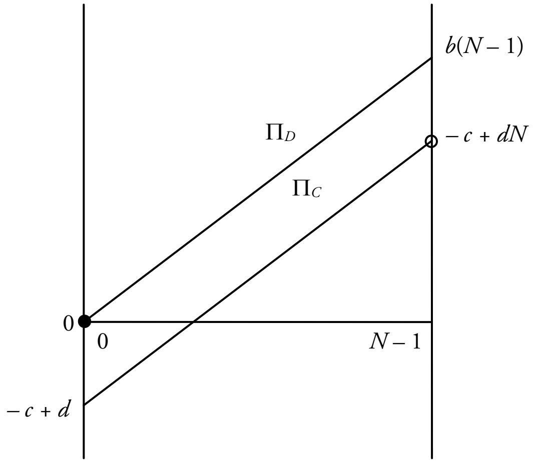
The third feature of the prisoner's dilemma is that the Nash equilibrium of the one-shot game is inefficient; all N countries would prefer an alternative feasible outcome where at least some countries play Cooperate to the Nash equilibrium in which no country plays Cooperate. I shall strengthen this assumption slightly and assume that the aggregate payoff is strictly increasing in z (this will ensure that the aggregate payoff is maximized when all countries play Cooperate; that is, when $z = N$). A little calculus shows that this requires

$$-c + 2dz > b(2z - N) \quad \text{for all z, } 0 \leq z \leq N \tag{3}$$

If the gain to any country i of one more of the other countries playing Cooperate is the same, irrespective of whether i plays Cooperate or Defect, then $b = d$. This situation is illustrated in Figure 1 (see also Schelling, 1978). If, however, the gain to any country i of one more of the other countries playing Cooperate is greater if i plays Cooperate also, then $d > b$. In this case, cooperation would exhibit a kind of increasing returns. I allow for both cases and so assume $d \geq b$.
To sum up, in addition to (1), (2), and (3), the model also assumes:

$$c > d \geq b > 0 \tag{4}$$

With this formulation, the equilibrium of the one-shot, N-player prisoner's dilemma game is unique: all countries play Defect in equilibrium. This equilibrium is inefficient: every country strictly prefers the outcome in which all countries play Cooperate. The latter outcome, called the full cooperative outcome, maximizes the aggregate welfare of all countries. The problem of international cooperation,



Number of other countries that play Cooperate

Figure 1. The $N \times 2$ (Symmetric) Prisoner's Dilemma Game ($b = d$)

at least as defined here, is to sustain the latter outcome as an equilibrium of a repeated game by means of a strategy of reciprocity.

Notice that I have defined the international cooperation problem as one where no country can be excluded from enjoying the benefits associated with cooperation by others. The problem of sustaining international cooperation is thus defined here as a problem of providing an international public good. Protection of the ozone layer and climate change mitigation are examples of global public goods. Other problems of interest are not suited to the model constructed here – international trade agreements being only one example.

3. Individual Rationality in the Infinitely Repeated, *N*-Player Prisoner's Dilemma

Suppose that the one-shot game is repeated infinitely often and that, against this background, the N players negotiate an agreement in which they all pledge to play the famous Grim strategy; that is, they all agree to play Cooperate in period 0 and to play Cooperate in every subsequent period provided no player ever played Defect in the past but that, should Defect ever be played by any player, every player must thereafter play Defect forever.

Grim has two attractive features. The first is that play Grim is a Nash equilibrium: given that the other players play Grim, any player j can do no better than to play Grim. To see that this is so in the present model, suppose player j deviates in period t. It will then get a payoff of $\Pi_D(N-1) = b(N-1)$ at time t. By (2) we know that $\Pi_D(N-1) > \Pi_C(N)$. So j gains initially from the defection. However, j will lose in the long run if the threatened punishment really is carried out. To know whether j can gain on balance from defecting, we need only compare the per-period payoff in the cooperative and punishment phases, assuming that the rate of discount is negligibly small. In the punishment phase, j gets an average payoff of $\Pi_D(0) = 0$. In the cooperative phase, j gets a per-payoff of $\Pi_C(N) = -c + dN$. Inequality (3) tells us that the latter payoff exceeds the former (since (3) must hold for $z = N/2$). So no player can gain by deviating unilaterally from Grim in a cooperative phase.

The Nash equilibrium is a rather weak requirement. For it is reasonable to ask: if a country did deviate, would every country really play Grim? Suppose that every other country plays Grim in a punishment phase. Will country i want to play Grim also? If i plays Grim, it will get a per-period payoff of $\Pi_D(0) = 0$. If i deviates, it will get a per-period payoff of $\Pi_C(1) = -c + d$. By (2), the former payoff exceeds the latter. So the threat to implement the Grim punishment is individually rational. Furthermore, this is true for any N.

It is of course true that *any* feasible, individually rational outcome of the one-shot game can be sustained as a subgame perfect equilibrium of the infinitely repeated game provided the players are sufficiently patient (see, for example, Fudenberg and Maskin, 1986). For example, the strategy Always Play Defect sustains the equilibrium of the one-shot game as a subgame perfect equilibrium of the repeated prisoner's dilemma. But given that, in the context of international negotiations, the players are able to meet, to deliberate openly on their predicament, to negotiate, it would be collectively irrational for them to choose to sustain a pareto-inefficient outcome from the set of all outcomes that can be supported as subgame perfect equilibria. So while the one-shot game cannot explain how countries could *ever* cooperate, the infinitely repeated game cannot explain why countries do not *always* cooperate. Theories built on either edifice will thus lack any cutting power; they will not be able to make sharp predictions.

It might seem from this discussion that the assumption of collective rationality favors cooperation.[10] I show later, however, that this is not so. More than that, I show that this assumption gives the cutting power that we desire in a theory.

4. Collective Rationality in the Infinitely Repeated, *N*-Player Prisoner's Dilemma

Though Grim is subgame perfect, it *seems* incredible because it is grossly unforgiving. Indeed, it is precisely for this reason that the famous Tit-for-Tat strategy appeals more to our intuition. But Tit-for-Tat is *not* subgame perfect; it is not an individually rational strategy. If a party deviates and then reverts to Tit-for-Tat, and if all other players play Tit-for-Tat, then the one-off defection results in an 'unending echo of alternating defections' (Axelrod, 1984: 176). In other words, the players could do better by deviating from Tit-for-Tat after the one-off deviation has occurred.

Contrary to intuition, Grim can can claim to be superior to Tit-for-Tat. But there is a problem with Grim that individual rationality fails to reveal. As our intuition suggests, Grim is too unforgiving. Though countries do not have an incentive to deviate from Grim unilaterally, they do have an incentive to deviate *en masse*. Grim is not a collectively rational strategy.

To see this, consider the $N = 2$ game and suppose that one of these countries, country j, deviates from Grim. Then each player will get an average per-period payoff in the punishment phase of 0. Though neither player can do better by deviating in the punishment phase, both players would do better collectively by renegotiating their agreement and restarting a cooperative phase, for they would then each get an average payoff of $-c + 2d$, and by (3) we know that $-c + 2d > 0$.

[10] Indeed, were I to drop the assumption of individual rationality, collective rationality would sustain *only* the full cooperative outcome.

Moreover, consistency demands that the theory allow them to renegotiate. The folk theorems are intended to explain how cooperation might emerge as an equilibrium, but they only allow players to begin a cooperative phase once (usually, in some period labelled 0). This is arbitrary. The theory should also allow cooperation to restart following a period of defection. Put differently, the theory should acknowledge that the players cannot make a credible commitment not to renegotiate. A self-enforcing treaty must not only be subgame perfect but also immune to renegotiation.[11]

A strategy that satisfies these requirements is a close cousin of Tit-for-Tat, Getting-Even.[12] This requires that country i play Cooperate unless i has played Defect less often than any of the other players in the past. The main difference between Tit-for-Tat and Getting-Even is that the latter strategy imposes a punishment that is more proportionate to the harm caused by the deviation. In a two-player game, if one player deviates for 20 periods and then reverts to cooperation, Tit-for-Tat demands that the other player revert to cooperation immediately after the first player has done so. Getting-Even, by contrast, requires the other player not to revert to cooperation for 20 periods.

To show that Getting-Even is both individually and collectively rational, consider again the N-player game. Suppose j deviates at time t and then reverts to Getting-Even in period $t + 1$. j then gets a payoff of $b(N - 1)$ in period t, a payoff of $- c + d$ in the punishment period, and a per-period payoff of $- c + dN$ from period $t + 2$ onwards. Had j not deviated, it would have gotten a payoff of $- c + dN$ every period from time t onwards. Since we are taking discount rates to be vanishingly small, deviating is individually irrational provided j would get a larger total payoff in periods $t + 1$ and $t + 2$ by playing Cooperate than by playing Defect. If j does not defect, it will get $2(- c + dN)$ in these periods. If j does defect and then reverts to Getting-Even, it will get $b(N - 1) - c + d$ in these periods. Play Getting-Even is thus individually rational if $2(- c + dN) > b(N - 1) - c + d$ or $- c + 2dN - bN > d - b$. Setting $z = N - 1$, (3) implies $- c + 2dN - bN > 2(d - b)$. So, provided $d \geq b$, (3) implies that Getting-Even is an equilibrium strategy. Setting $z = N$, (3) implies $- c + 2dN - bN > 0$. So Getting-Even is also an equilibrium strategy for $d < b$.

However, Getting-Even is only subgame perfect provided $d \geq b$. To see this,

[11] It might be argued that it should also not be possible for any coalition of countries, taking the actions of all others as given, to agree to deviate from the agreement; that it should not be possible for any subcoalition to agree to deviate from this alternative agreement; and so on. In other words, it might be argued that treaties should be coalition-proof Nash equilibria (see Bernheim et al., 1987). However, application of this concept to the infinitely repeated prisoner's dilemma poses certain technical problems, as noted by Bernheim et al. (1987).

[12] The concept of a renegotiation-proof equilibrium used here is due to Farrell and Maskin (1989). Van Damme (1989) derives the strategy which supports full cooperation as a renegotiation-proof equilibrium of the two-player prisoner's dilemma. See also Myerson (1991), who gave this strategy the name, 'Getting-Even.' My contribution here is to extend the use of this concept to the $N > 2$ case and to apply it to international cooperation problems.

suppose j deviates at time t and then reverts to Getting-Even. In period $t + 1$, j therefore plays Cooperate, while all other players play Defect. Any player i, $i \neq j$, gets a payoff of b in period $t + 1$ and a payoff of $-c + dN$ in every subsequent period if all players play Getting-Even from period $t + 1$ onwards. If i deviates in period $t + 1$ and then reverts to Getting-Even in period $t + 2$, however, it gets a payoff of $- c + 2d$ in period $t + 1$ and a payoff of $b(N - 1)$ in period $t + 2$; thereafter, i gets $- c + dN$ every period. Deviating is therefore irrational for i provided $b - c + dN \geq - c + 2d + b(N - 1)$ or $d \geq b$. This last requirement holds by (4).

To sum up so far: like Grim, Getting-Even is individually rational. I now show that, unlike Grim, Getting-Even is also collectively rational.

Getting-Even will be collectively rational if all countries have no incentive to renegotiate the agreement. If every country other than j plays Getting-Even in a punishment phase, after j has reverted to Getting-Even, then they will each get a payoff of b per period. If they deviate *en masse*, however, then they will each get $- c + dN$ per period. It will thus not be in their collective interests to deviate if

$$(b + c)/d \geq N. \qquad (5)$$

Since $d \geq b$ by assumption, (5) implies that $(d + c)/d \geq N$, and this in turn implies that all the countries called upon to punish j for cheating cannot do better collectively than to play Defect in the punishment phase, even if j plays Defect in this phase also. Agreements that satisfy (5) are not vulnerable to renegotiation. The threats needed to sustain full cooperation in these agreements are credible.

To sum up: I have shown that Getting-Even can sustain full cooperation by means of a self-enforcing agreement if (5) holds. I have *not* shown that there does not exist an alternative strategy that can do better than Getting-Even (that is, a strategy that can sustain full cooperation using a weaker punishment, and so allow full cooperation to be sustained for a larger N). However, in the Appendix I show that Getting-Even cannot be bettered, as long as we hold on to the assumptions of individual and collective rationality. Result (5) is robust.

Inequality (5) tells us that the full coooperative outcome can only be sustained as an equilibrium of the repeated game if N is not ‘too large’. Notice that, since (2) must hold for $z = 1$, $(b + c)/d < 2$. So we know that the full cooperative outcome of the generic 2 × 2 prisoner’s dilemma game can be sustained as an equilibrium of the repeated game. This is not a new result (see van Damme, 1989; Myerson, 1991), but (5) shows just how special the two-player game is. It may not be possible for even three countries to sustain the full cooperative outcome by means of a self-enforcing agreement.

Importantly, (5) tells us that the maximimal value of N that can sustain the full cooperative outcome as an equilibrium is not fixed but depends on the parameter values. Consider some examples. Suppose $b = d = 3$ and $c = 4$. Then (2) and (3) hold for $N \geq 2$, but at most two countries can sustain the full cooperative outcome as an equilibrium of the repeated game. Suppose instead that $b = 2$, $d = 3$, and $c = 10$. Then (2) and (3) hold for $N = 6$ and $N = 7$, but (5) says that at most four countries can sustain full cooperation by means of a self-enforcing agreement. Finally, suppose $b = d = 1$, and $c = 149$. Then (2) and (3) hold for $N \geq 150$ while full cooperation can be sustained as an equilibrium only so long as $N \leq 150$. Keohane (1984) has argued that, for international relations problems, the number of players is 'small,' even in the case of global negotiations (in 1984, when Keohane made this argument, there were about 150 countries in the world; today there are almost 200). But the theory developed here shows that whether the international system *is* 'small' depends on the nature of the cooperation problem.

More than this, the theory implies that the number of countries in the world is 'small' only with regard to issues for which the total gains to cooperation are 'small.' In other words, when cooperation is needed most, the international system is least capable of sustaining cooperation by a self-enforcing agreement. To see this, notice that the gains to cooperation are $N[\Pi_C(N) - \Pi_D(0)] = N(-c + dN)$. The gains to cooperation are thus decreasing in c and increasing in d. But from inequality (4) we know that the maximal value of N that can sustain full cooperation as an equilibrium is increasing in c and decreasing in d. So the international system can only sustain full cooperation among *all* countries when the gains to cooperation are 'small.'

Does this result speak to any real world problems? I have shown elsewhere (Barrett, 1999) that the aggregate gains to cooperation are small in the case of stratospheric ozone depletion. This is not because the world would not benefit from a ban on ozone-depleting substances. To the contrary, the reason is that the benefit of a ban is so large relative to the cost, that every industrial country would want to ban these chemicals unilaterally, even if no other country did so. The challenge to the Montreal Protocol was to make it attractive for poorer countries also to ban these substances, and for the ban by signatories to be made effective by ensuring that production would not relocate to nonsignatory countries.[13]

5. Compliance Enforcement and Free-Rider Deterrence

The theory outlined here teaches that cooperation can be sustained by a self-enforcing treaty which incorporates a strategy of reciprocity. But Chayes and Chayes (1991: 313) observe that 'not only are formal enforcement mechanisms seldom used to secure compliance with treaties, but they are rarely even embodied in the treaty text'. Now, the fact that such enforcement mechanisms are seldom

[13] The former problem requires the use of 'carrots' or side payments. For an analysis of how carrots can aid cooperation, see Barrett (1998). The latter problem is sometimes called 'trade leakage,' and is discussed in Barrett (1997).

used is entirely consistent with the theory developed here. In equilibrium, no party would deviate from the treaty because the threat to carry out the punishment is credible. Where the theory and practice of international coooperation seem to clash is in the observation that compliance enforcement mechanisms are rarely expressed in black and white. The reason may be that the theory is wrong and such mechanisms are not needed, as the Chayes's argue; or it may be that, as Downs et al. (1996) maintain, the theory is right and the fact that such mechanisms are not incorporated in treaties implies that agreements typically do not improve much on the noncooperative outcome.

To illuminate this debate, I distinguish between free-rider deterrence and compliance enforcement by representing international cooperation as a stage game: in Stage 1, countries choose whether to be signatories or nonsignatories to an international agreement; in Stage 2, signatories choose *jointly* whether to play Cooperate or Defect; and in Stage 3, nonsignatories choose independently whether to play Cooperate or Defect. I assume that the compliance norm has been fully internalized, so that all signatories comply with the obligations they negotiate in Stage 3. As noted in the Introduction, this assumption is merely a tactic. I use it to see whether internalization of the compliance norm matters.

As usual, the equilibrium is found by solving the stage game backwards. Assuming that all actions are publicly observable, the strategies of each player will generally be contingent on the history of the game. However, the stage game version of the prisoners' dilemma is special in that the history of the game is irrelevant to nonsignatories; for them, play Defect is a dominant strategy. If signatories were to choose whether to play Cooperate or Defect independently, then they too would play Defect. However, signatories to a treaty do not choose their actions independently. They negotiate their choice of actions and it would be collectively irrational for them to put their signatures on a treaty that did not maximize their joint payoff.

Let k denote the number of signatories, and let signatories be identified by the subscript s and nonsignatories by the subscript n. Then, for the two-player game, if $k = 1$ the sole signatory will play Defect and get a payoff $\Pi_s = 0$ (if this country played Cooperate instead it would get a payoff of $-c + d$, and by (2), $-c + d < 0$), while if $k = 2$ both signatories will play Cooperate (since $-c + 2d > (b - c + d)/2$ by (3)) and get a payoff $\Pi_s = -c + 2d$ each. Nonsignatories can do no better than to play Defect, whatever signatories do, and so they get a payoff $\Pi_n = 0$ if $k = 0$ or $k = 1$.

These payoffs can be worked out by each country before the Stage 1 game is played. So in Stage 1, each country will know the consequence of choosing to be a signatory or nonsignatory, taking as given the choice of the other country to be a signatory or nonsignatory. Assuming that a country will accede to a treaty if, in doing so, it is not made worse off, there is a unique equilibrium. It is that both countries are signatories and that both play Cooperate. The institution of the treaty coupled with the compliance norm thus transforms the dilemma game into one in which full cooperation is sustained as an equilibrium.

But full cooperation will not always be sustained as an equilibrium of the transformed game. Suppose the payoff functions are given by $\Pi_D = 3z$ and $\Pi_C = -4 + 3z$. Then we obtain the above result for $N = 2$. Not so if $N = 5$. For the transformed game, nonsignatories will play Defect in equilibrium. If there is only one signatory, it too will play Defect (if this country plays Defect it gets $\Pi_S = 0$; if it plays Cooperate it gets $\Pi_S = -1$). However, if there are two or more signatories, they will each get a higher payoff if they both play Cooperate (for example, if $k = 2$, each signatory gets $\Pi_S = 0$ if they both play Defect and $\Pi_S = 2$ if they both play Cooperate). And so on. It can be shown that, in equilibrium, $k^* = 2$ signatories play Cooperate and $N - k^* = 3$ nonsignatories play Defect. The full cooperative outcome is not sustained as an equilibrium of this game, even though the compliance norm is assumed to have been fully internalized.

To generalize even further, suppose the payoff functions for the N-player dilemma game are given by equations 1. Then signatories will play Cooperate provided the payoff they each get by playing Cooperate exceeds the payoff they each get by playing Defect, or $k > c/d$; otherwise, signatories can do no better collectively than to play Defect. Because play Cooperate is not an equilibrium of the one-shot prisoner's dilemma game, we know that $c/d > 1$ and so $k^* \geq 2$. As in the repeated game model, cooperation can always be sustained as an equilibrium for the special two-player case.

Since, by assumption, full cooperation requires that all players play Cooperate, it must be true that $N > c/d$. Let k^0 be the smallest integer greater than c/d. Then we know that $k^* \geq k^0$. But when $k = k^0$, no nonsignatory would wish to accede to the treaty. To see this, notice that, if $k = k^0$, a nonsignatory gains by acceding to the treaty if $(d - b)\, k^0 > c - d$. But, by (2), $(d - b)z < c - d$ for all z, $0 \leq z \leq N - 1$. This is a contradiction. Once there are k^0 signatories, it would be irrational for another country to accede to the treaty. Hence, the equilibrium number of signatories must be $k^* = k^0$ (assuming that the solution is 'interior'). Figure 2(a) illustrates the solution for $k^* < N$ and Figure 2(b) for the case where $k^* = N$.

Full cooperation can only be sustained as an equilibrium of this transformed game if signatories can do no better collectively than to play Defect when $k = N - 1$ and to play Cooperate only when $k = N$. The latter requirement holds by (3). The former holds provided $0 \geq -c + d(N - 1)$ or

$$(d + c)/d \geq N. \qquad (6)$$

Notice that (6) can be interpreted as saying that an agreement to play Cooperate would only come into force (that is, would only be legally binding on

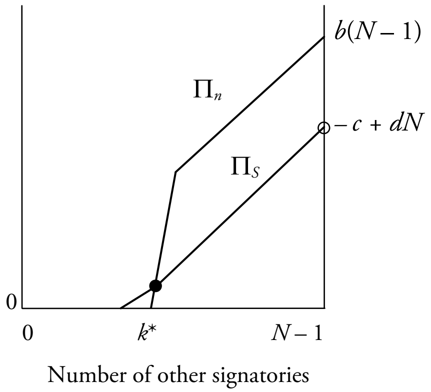


Figure 2a. The N × 2 (Symmetric) Transformed Prisoner's Dilemma Game, k < N*

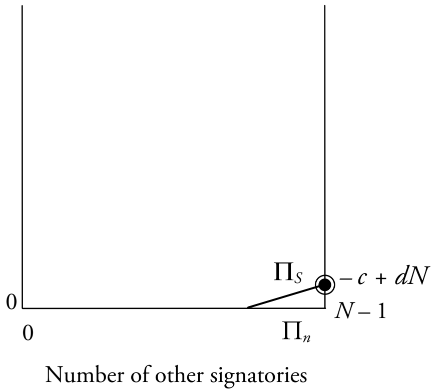


Figure 2b. The N × 2 (Symmetric) Transformed Prisoner's Dilemma Game, k = N*

the countries that had ratified it) if all N countries have ratified it. Hence, k^* can be interpreted as the minimum participation level prescribed by international treaties. Of course, the case where $k^* = N$ is special. And it is a feature of most treaties that the actual number of parties usually exceeds the number prescribed by the minimum participation clause. This suggests that in the majority of treaties the minimum participation clause may serve as a coordination device rather than as a mechanism for deterring free-riding.[14]

Upon comparing (5) and (6) one finds that, if $b = d$, then the maximum number of countries that can sustain the full cooperative outcome as a self-enforcing agreement will be the same for both models. If, however, $d > b$ – if there are increasing returns to cooperation – then a smaller number of countries can sustain the full cooperative outcome as an equilibrium in the repeated game model as compared to the stage game model. However, as noted in the introduction, this does not mean that the assumption of full compliance buys any additional cooperation. The stage game model is essentially one-shot; it does not allow for reactions, and so it cannot describe fully an appropriate strategy of reciprocity.

The main reason for using the stage game model is to show that the vital qualitative insight of the repeated game model holds here as well. Recall that the total gain to cooperation, $N(-c + dN)$, is decreasing in c and increasing in d. By contrast, k^* is increasing in c and decreasing in d (ignoring the integer problem). This means that, for N given, k^* will tend to be 'large' ('small') when the total gain to cooperation is 'small' ('large'). The international system is able to sustain less cooperation the greater is the potential gain to cooperation – that is, the greater is the need for cooperation (see also Barrett, 1994).

[14] See Barrett (1997), where the minimum participation clause actually emerges as an equilibrium.

Notice that, in equilibrium, nonsignatories get a higher payoff than signatories. Nonsignatories (of which there are $N - k^*$) free-ride. The underlying game of whether to play Cooperate or Defect is a prisoner's dilemma game, but the transformed game of whether to be a signatory or nonsignatory to the treaty is a chicken game. Each country would prefer to free-ride, but if too few countries are parties to the treaty, it is in the interests of nonsignatories to accede. Though the players are symmetric by assumption, in equilibrium they behave differently. Some are signatories and play Cooperate; some are nonsignatories and play Defect. The model cannot identify which countries will be signatories and which nonsignatories (though the identities of these countries can be determined if countries make their stage 1 choices in sequence; the first $N - k^*$ countries to choose will all choose not to be signatories and the last k^* to choose will all choose to be signatories), but as the countries are symmetric this does not matter.[15]

The essential lesson of the stage game is that, despite the assumption of full compliance, a self-enforcing treaty may only be capable of sustaining $k^* < N$ signatories. Free-riding may be a problem for international cooperation, even if compliance is not. At the very least, sticks are needed to deter free-riding, though the constraints on individual and collective behavior may be such that the full cooperative outcome cannot be sustained by international treaty. Large sticks may be needed to deter free-riding but large sticks may not be credible.

Though I am unable to settle the dispute about compliance, the theory developed here does broaden the debate. It suggests that, even if the Chayes's are right that compliance is not a problem, they may be wrong that sanctions are not needed to sustain cooperation or that the international system sustains anything like full cooperation. It suggests too that Downs et al. may be right that full cooperation typically has not been sustained, but that they may be wrong in implying that the reason for this is weak enforcement. Free-rider deterrence may be the greater problem.

What constrains cooperation in the stage game, as in the repeated game, is the assumption that signatories negotiate a collectively rational agreement. If we drop this requirement, then the assumption that the compliance norm has been internalized will ensure that full cooperation can always be sustained as an equilibrium. For if signatories could be sure of complying with *any* agreement, then to sustain full cooperation as an equilibrium would only require an agreement which says that each country will play Cooperate provided all others do, but that, should any other country play Defect instead, then all the other countries will punish this defection. In general, however, such an agreement will not be collectively rational. Should one country play Defect, it will not generally be collectively rational for the remaining $N - 1$ countries to punish the deviation.

[15] This will not be true when countries are strongly asymmetric; see Barrett (1998).

6. Conclusions

The central idea behind the theory presented here is that the institutions that sustain international cooperation must be both individually and collectively rational: individually rational because the international system is anarchic; collectively rational because countries cooperate explicitly and can renegotiate their treaties at any time. When combined, these requirements give the theory of international cooperation great cutting power. The theory predicts that the full cooperative outcome of the N-player, prisoner's dilemma can only be sustained by a self-enforcing treaty when N is 'small.' For global problems (that is, problems for which N is 'large'), the theory predicts that full cooperation can only be sustained by a self-enforcing treaty when the gains to cooperation are 'small.'

These are powerful if depressing predictions. They are not, however, context-free. In an environment richer than the one analysed here, it is possible that more cooperation could be sustained by a self-enforcing treaty. For example, I have shown elsewhere (Barrett, 1997) how linking the provision of a global public good to international trade allows the space of punishment strategies to be expanded. The credible threat of trade sanctions *may* be able to sustain cooperation where the threat to withdraw provision of a public good cannot. In fact, it is by the threat of imposing trade sanctions that free-riding has been deterred in the Montreal Protocol. Moreover, the threat of trade sanctions has also helped to enforce compliance with the agreement. (That sanctions for non-compliance should be linked to sanctions for non-participation is, of course, a conclusion of this paper.) I have also shown how side payments can help to increase participation in an agreement when countries are strongly asymmetric (Barrett, 1998).[16] But even where the strategy space can be expanded in these ways, the twin requirements of individual and collective rationality may prevent countries from sustaining full cooperation. Certainly, there should be no presumption that the international system, attached as it is to the principle of sovereignty, is always capable of sustaining full cooperation. *That* conclusion, however unwelcome, does seem robust.

Appendix

Getting-Even, as defined in this paper, assumes that, were j to deviate, then *all* the $N - 1$ other countries must play Defect in a punishment phase. In doing so, these countries harm themselves as well as j, and this is what makes sustaining full cooperation more difficult as N increases. So the question arises: can an alternative strategy – one that harms the $N - 1$ other countries less – sustain full cooperation?

[16] Of course, one feature of asymmetry is that the failure to cooperate matters less, and so the aggregate gains to cooperation are smaller for a given N compared with the symmetric case. This was perhaps first shown by Olson (1965). A formal demonstration can be found in Barrett (1998).

This will not be possible for $N = 2$, because obviously j must be punished for deviating and when $N = 2$ there is only one other country that can do so. However, it is not obvious that, when $N > 2$, all the other $N - 1$ countries must play Defect in a punishment phase. Let us then suppose that m of the $N - 1$ other countries play Defect in the punishment phase (so that $N - m - 1$ of the $N - 1$ other countries play Cooperate in the punishment phase). Call this the m-Getting-Even strategy.

Full cooperation can be sustained as an equilibrium if two conditions are satisfied. First, we require that j cannot do better than to play m-Getting-Even given that every other country does so; that is, we require

$$\max\,(b(N - m - 1) - c + d(N - m)) \leq - c + dN \qquad \text{(A.1)}$$

By (2), $b(N - m - 1) > - c + d(N - m)$. So (A.1) implies

$$b(N - m - 1) \leq - c + dN \qquad \text{(A.2)}$$

We also require that each of the $N - 1$ other players cannot do better than to play m-Getting-Even in a punishment phase. That is, we require

$$b(N - m) \geq - c + dN \qquad \text{(A.3)}$$

for the m countries that play Defect in the punishment phase and

$$- c + d(N - m) \geq - c + dN \qquad \text{(A.4)}$$

for the $N - m - 1$ other countries that play Cooperate in the punishment phase. But (A.4) reduces to $- dm \geq 0$, implying that we must have $m = 0$. Of course, if $m = 0$ – if none of the $N - 1$ other countries plays Defect in a punishment phase – then j will not be punished. So the m-Getting-Even strategy cannot sustain full cooperation as an equilibrium, except for the special case where $m = N - 1$ (for in this case, (A.4) drops out and (A.3) reduces to (5)), provided we require that *all* the $N - 1$ countries do not want to renegotiate the agreement.

Now, it might be argued that this requirement is overly strong. Suppose we allow transfers between the $N - 1$ countries called upon to punish j. Then renegotiation will be prevented if the $N - 1$ other countries receive *on average* at least as large a payoff when implementing the strategy as when reverting to full cooperation. However, collective rationality will in this case require that the $N - 1$ other countries choose m so as to maximize their aggregate payoff in the punishment phase. That is, instead of (A.3) and (A.4) we require

$$\max_m \{mb(N-m) + (N-m-1)[-c + d(N-m)]\} \geq (N-1)(-c + dN) \tag{A.5}$$

Solving the left-hand side of (A.5), the first-order condition requires

$$b(N-2m) + c - d[2(N-m) - 1] = 0 \tag{A.6}$$

The second-order conditions for a maximum require $2(d-b) < 0$. However, by assumption, $d \geq b$. Hence, the solution to the maximization problem must lie at a corner; (A.5) will require either $m = 0$ or $m = N - 1$.

Of course, (A.2) must hold, and this implies

$$m \geq [b(N-1) - (-c + dN)]/b \tag{A.7}$$

By (2), the numerator on the right-hand side of (A.7) is positive. So the solution must require $m > 0$. $m = N - 1$ will be the solution to the left-hand side of (A.5) if the aggregate payoff of the $N - 1$ other countries is at least as high when $m = N - 1$ as when $m = 0$. Upon substituting, we require

$$b \geq -c + dN \tag{A.8}$$

But this is the same as (5). Hence, there does not exist an alternative individually and collectively rational strategy that can improve on Getting-Even.

References

Axelrod, R. (1984), *The Evolution of Cooperation*, New York: Basic Books.

Axelrod, R. and R.O. Keohane (1985), ‘Achieving cooperation under anarchy: Strategies and institutions’, *World Politics* **38**, 226–254.

Barrett, S. (1994), ‘Self-enforcing international environmental agreements’, *Oxford Economic Papers* **46**, 878–894.

Barrett, S. (1997), ‘The strategy of trade sanctions in international environmental agreements’, *Resource and Energy Economics* **19**, 345–361.

Barrett, S. (1998), ‘Cooperation for sale’, mimeo, London Business School.

Barrett, S. (1999), ‘Montreal versus Kyoto: International cooperation and the global environment’, in I. Kaul, I. Grunberg and M.A. Stern (eds), *Global Public Goods: International Cooperation in the 21st Century*, New York: Oxford University Press.

Bernheim, B.D., B. Peleg and M.D. Whinston (1987), ‘Coalition-proof Nash equilibria. I. Concepts’, *Journal of Economic Theory* **42**, 1–12.

Chayes, A. and A.H. Chayes (1991), ‘Compliance without enforcement: State regulatory behavior under regulatory treaties’, *Negotiation Journal* **7**, 311–331.

Chayes, A. and A.H. Chayes (1995), *The New Sovereignty*, Cambridge, MA: Harvard University Press.

Downs, G.W., D.M. Rocke and P.N. Barsoon (1996), ‘Is the good news about compliance good news about cooperation?’, *International Organization* **50**, 379–406.

Farrell, J. and E. Maskin (1989), 'Renegotiation in repeated games', *Games and Economic Behavior* **1**, 327–360.

Fudenberg, D. and E. Maskin (1986), 'The Folk theorem in repeated games with discounting or with incomplete information', *Econometrica* **54**, 533–556.

Kandori, M. (1992), 'Social norms and community enforcement', *Review of Economic Studies* **59**, 63–80.

Keohane, R.O. (1984), *After Hegemony*, Princeton, NJ: Princeton University.

Keohane, R.O. (1986), 'Reciprocity in international relations', *International Organization* **40**, 1–27.

Keohane, R.O. and E. Ostrom (1994), 'Introduction', *Journal of Theoretical Politics* **6**, 403–428.

Myerson, R.B. (1991), *Game Theory: Analysis of Conflict*, Cambridge, MA: Harvard University Press.

Olson, M. (1965), *The Logic of Collective Action,* Cambridge, MA: Harvard University Press.

Ostrom, E. (1990), *Governing the Commons*, Cambridge: Cambridge University Press.

Schelling, T.C. (1978), *Micromotives and Macrobehavior*, New York: W.W. Norton & Co.

Snidal, D. (1994), 'The politics of scope: Endogenous actors, heterogeneity and institutions', *Journal of Theoretical Politics* **6**, 449–472.

van Damme, E. (1989), 'Renegotiation-proof equilibria in repeated prisoner's dilemma', *Journal of Economic Theory* **47**, 206–17.

Young, O.R. and G. Osherenko (eds.) (1993), *Polar Politics*, Ithaca, NY: Cornell University Press.

Bidding for the Surplus: A Non-Cooperative Approach to the Shapley Value

David Pérez-Castrillo and David Wettstein

Paper presented at the CTN Workshop 2000. Reprinted from *Journal of Economic Theory,* **100**(2), October 2001, 274–294, with permission from Elsevier.

We propose a simple mechanism to determine how the surplus generated by cooperation is to be shared in zero-monotonic environments with transferable utility. The mechanism consists of a bidding stage followed by a proposal stage. We show that the subgame perfect equilibrium outcomes of this mechanism coincide with the vector of the Shapley value payoffs. We extend our results to implement the weighted Shapley values. Finally, we generalize our mechanism to handle arbitrary transferable utility environments. The modified mechanism generates an efficient coalition structure, and implements the Shapley values of the super-additive cover of the environment.

1. Introduction

The Shapley value has long been a central solution concept in cooperative game theory. It was introduced in Shapley (1953) and was seen as a reasonable way of distributing the gains of cooperation among the players in the game. It is the most studied and widely used single-valued solution concept in cooperative game theory. It has generated various axiomatizations that demonstrate its fairness and consistency properties (see, for instance, Myerson (1980), and Hart and Mas-Colell (1989), and has been used to impute costs and benefits as in cases of airport runways, phone networks, and political situations.[1]

We wish to thank Sergiu Hart for several helpful suggestions and conversations, and Ezra Einy, Gerard Hamiache and Roberto Serrano as well as participants in seminars at several universities for helpful comments. Pérez-Castrillo acknowledges financial support from the DGES PB 97-0181 and SGR 96-75. Part of this research was conducted while Wettstein was visiting the Universitat Autònoma de Barcelona, with a grant from the Generalitat de Catalunya, and both authors were visiting the University of Copenhagen, whose financial support is acknowledged.

[1] For a nice introduction to the Shapley value and, in particular, its applications, see, for example Roth (1988).

A natural question concerning the Shapley value is whether the agents can reach it through non-cooperative behavior. In other words, is it possible to find a non-cooperative framework that gives rise to the Shapley value as the result of equilibrium behavior? This is part of the Nash program, which tries to provide a non-cooperative foundation for cooperative solution concepts. Several papers have addressed this question in different ways. We will comment on them later in this introduction.

In this paper, we provide a simple non-cooperative game whose outcome always coincides with the Shapley value for zero-monotonic games in characteristic form. We call this game the '*bidding mechanism*'. The basic idea of the bidding mechanism is quite simple. We let one of the players make a proposal to each of the other players, a proposal that will either be accepted by all the other players (in which case the proposal becomes final) or rejected. In the latter case, the proposer is now on his own and the rest of the players play the same game again. If the proposal is accepted, the proposer can form the grand coalition of all the players and collect the value generated in exchange for the proposed payments to the rest of the players.

The question of how the proposer is determined is, of course, central to the design of the bidding mechanism. Indeed, in some games, being the proposer could prove to be beneficial, while in other games it is preferable not to be the proposer. Hence, before the proposal stage is reached, the players will bid to become the proposer, where bids can be positive or negative.[2] The player with the highest 'net bid' (the difference between the sum of the bids he makes to the others minus the sum of the bids the others make to him) becomes the proposer and, before proceeding to the proposal stage, pays the bids to the other players. We will show that in the subgame perfect equilibria (SPE) of the bidding mechanism a proposer is determined who will make a proposal that will be accepted by the others. For the proposer, the difference between the value of the grand coalition and the payments and bids paid is her Shapley value. For each of the other players as well, the sum of the bid received plus the accepted proposal is his Shapley value.[3]

Several features of our game make it attractive and different from previous non-cooperative approaches to the Shapley value. First, the players obtain the Shapley value in *every* equilibrium outcome of the game; that is, the implementation is not in expected terms. Also, the game does not imply any *a priori* randomization that imposes some order on the moves of the players. By

[2] Crawford (1979) also made use of a bidding stage in a procedure to generate Pareto-efficient egalitarian-equivalent allocations. The discrete time non-cooperative coalitional bargaining game proposed by Evans (1997) to implement the core in subgame perfect equilibria also introduced simple bidding by the players for the right to make an offer.

[3] The equilibrium strategies are unique if the game is strictly zero-monotonic. Otherwise, there might be other equilibria in addition to this one, but they still yield the Shapley value.

adjusting his bids, every player can determine whether he or someone else will be the proposer. Second, the rules of the game are very natural and do not rely on 'random' meetings or probabilities that are close to the actual definition of the Shapley value. Hence, the implementation is less 'obvious', and provides further support for the use of the Shapley value. Third, the game is finite. Moreover, at equilibrium, it ends in one stage if the game is strictly zero-monotonic (a stage includes three periods of play: bidding, proposing, and accepting or rejecting). Fourth, the strategies played by the players at equilibrium are simple and intuitive. Furthermore, even though the Shapley value plays no role in specifying the rules of the game, the equilibrium strategies are intimately related to the Shapley value itself.

Implementing the Shapley value is not straightforward. For example, Thomson (1988) focused on the problems created by strategic behavior and showed that an agent can obtain a better outcome by unilaterally misrepresenting his utility function. Several authors have attempted to realize the Shapley value and overcome such problems.

Gul (1989 and 1999) analyzed a transferable utility economy where random meetings between two agents occur. At each meeting, a randomly chosen party makes an offer to his partner. Acceptance of the offer means that the proposer buys the partner's resources. If the offer is rejected, the meeting dissolves and both agents stay in the market. For strictly convex games, as the time interval between meetings becomes arbitrarily small, the expected payoff of each player at an efficient stationary subgame perfect Nash equilibrium (SSPE) converges to his Shapley value. If strict convexity is replaced by strict superadditivity the convergence result holds for those efficient SSPE that entail immediate agreement (Gul, 1989 and Hart and Levy, 1999).

Evans (1996) showed that a simplified version of Gul's result follows from the following characterization of the Shapley value. Consider a cooperative game and an associated feasible payoff vector. Assume that players are randomly split into two groups and a representative player is chosen also at random from each group. These two players bargain with each other over how to split the total resources available to all of the players. Following the bargaining process each of the two players has to pay out of his share to the members of his group according to the pre-specified payoff vector. This procedure yields an expected payoff to any player that depends on the initial payoff vector, the random partition mechanism and the solution concept applied to two-person bargaining problems. The initial payoff vector is called consistent if it equals the expected payoff vector. If all partitions are equally likely and the bargaining solution splits the surplus equally, the Shapley payoff vector is the unique consistent payoff vector.

Hart and Mas-Colell (1996) proposed a different natural bargaining procedure that supports the Shapley value (as well as the Nash bargaining solution for pure bargaining problems). In their paper, the proposers are also chosen at random but the meetings are multilateral. Agreement requires unanimity. Disagreement puts the proposer in jeopardy, since there is a given probability that he may be removed from the game after a rejection. As the probability of removal becomes small, the SSPE of the procedure yield the Shapley value.[4] When the probability of removal is one, Hart and Mas-Colell (1996) as well as Mas-Colell (1988) showed that the expected payoff of any player coincides with his Shapley value. Their mechanism is then the same as our mechanism with the bidding stage replaced by a random determination of the proposer. Krishna and Serrano (1995) showed further that for removal probabilities close to one there is a unique SPE of the game proposed by Hart and Mas-Colell (1996) that yields the Shapley value payoff vector in expectation.

In a different spirit, Hart and Moore (1996) proposed a game in which agents are lined up and each agent makes an offer to the following agent, where the offer is a contract that may specify what offer this agent has to make to the agent after him. This game implements the Shapley value in SPE. Winter (1994) and Dasgupta and Chiu (1996) proposed demand commitment games in which each player can either make a demand to the following player or form a coalition satisfying the demands of some of the players preceding him. For strictly convex games, these mechanisms implement the Shapley value in SPE.[5] In these three works, the implementation is in expected terms since in the first stage of the game the order of the players (or the identity of the first player in Winter, 1994) is randomly chosen, each possible choice having the same probability.

A solution concept closely related to the Shapley value is the weighted Shapley value (Shapley, 1953). We also show that a very natural and simple modification of the bidding mechanism implements the weighted Shapley values.[6]

Finally, we generalize the bidding mechanism to deal with all transferable utility environments. In the *generalized bidding mechanism*, the proposer makes a proposal to each of the other players and, simultaneously, chooses the coalition she wants to form. If all the agents accept the proposal and the coalition, the coalition is formed, and the rest of the players proceed to play the same game among themselves (after having received the proposed payment by the proposer). In the case of rejection, the proposer is on her own and the remaining players play the same game again. In any SPE of this mechanism, the proposer makes a proposal that is accepted. The payoff of the proposer is the difference between the value of the coalition she formed and the payments and bids she made. The payoff to any player in the coalition is the sum of the bid and the proposal

[4] They also show that for NTU games, the limit of the SSPE (as the probability of removal becomes small) is the consistent value, a solution concept that was introduced by Maschler and Owen (1989, 1992).

[5] Winter (1994) also required either subgame consistency or strategic equilibria. Dasgupta and Chiu (1996) also developed an implementation for general games in characteristic form if there is an (external) planner who is able to impose a system of transfers and taxes.

[6] Hart and Mas-Colell (1989) also extended their results to weighted Shapley values.

accepted. The payoff to players outside the coalition is the sum of the bid, the proposal accepted, and their payment in the continuation game. Hence, the SPE of this mechanism determine a coalition structure and a sharing of the surplus generated under this particular structure. We show that at the SPE of the generalized bidding mechanism the players form an efficient coalition structure. Moreover, the final payments of the players coincide with the Shapley values of the super-additive cover of the game.[7]

The paper is organized as follows. Section 2 presents the basic cooperative definitions and Section 3 introduces the bidding mechanism and shows that it implements the Shapley value for zero-monotonic games. In Section 4 we extend our results by implementing the set of weighted Shapley values. In Section 5 we define the generalized bidding mechanism and show that it implements the Shapley value of the super-additive cover of the game. The paper concludes with a brief summary and discussion of further research.

2. The Cooperative Model

Consider a *cooperative game in characteristic form* (N, v), where $N = \{1,\ldots, n\}$ is the set of players and v: $2^N \rightarrow R$ is a characteristic function satisfying $v(\Phi) = 0$ where Φ is the empty set. For a coalition $S \subseteq N$, $v(S)$ represents the total payoff that the partners in S can jointly obtain if this coalition is formed. We say that the cooperative game (N, v) is *zero-monotonic* if $v(S) + v(\{i\}) \leq v(S \cup \{i\})$ for any subset $S \subseteq N$ with $i \notin S$. In a zero-monotonic game there are no negative externalities when a single player joins a coalition. In sections 2 to 4 of this paper, we restrict our analysis to zero-monotonic games.

We denote by $\phi(N) \in R^n$ the *Shapley value* of the cooperative game (N, v) which is defined by:[8]

$$\phi_i(N) = \sum_{S \subseteq N / \{i\}} \frac{|S|!\,(n-|S|-1)!}{n!}\left[v(S \cup \{i\}) - v(S)\right] \text{ for all } i \in N,$$

where $|S|$ denotes the cardinality of the subset S. The Shapley value can be interpreted as the expected marginal contribution made by a player to the value of a coalition, where the distributions of coalitions is such that any ordering of the players is equally likely. Also, Shapley (1953) characterized the value as the only function that satisfies symmetry, efficiency, a null player axiom, and additivity.

Given the cooperative game (N, v) and a subset $S \subseteq N$, we define the game

[7] If the game is super-additive, the grand coalition is an efficient structure and the Shapley value of the super-additive cover coincides with the Shapley value. Therefore, the final SPE outcome of the generalized bidding mechanism is the same as the final SPE outcome of the bidding mechanism.

[8] We use $\phi(N)$ instead of $\phi(N, v)$ for notational simplicity.

(S, v_S) by assigning the value $v_S(T) \equiv v(T)$ to every $T \subseteq S$. We write (S, v) instead of (S, v_S) for notational convenience. Similarly, $\phi(S) \in R^{|S|}$ denotes the Shapley value of the game (S, v).

3. The Bidding Mechanism

In this section, we design a non-cooperative game, which we call the *bidding mechanism*. The equilibrium outcomes of this mechanism always coincide with the Shapley value of the cooperative game (N, v) and thus this mechanism implements the Shapley value in subgame perfect equilibria. We propose a way to split the surplus of the cooperation that is based upon the idea that only one of the players will make a proposal to each of the other players. We invoke a notion of consistency in order to determine the outcome of the game if the proposal is rejected. Following a rejection the players other than the proposer play the same game again. Proceeding in this way, the payoff of an agent is sensitive not only to whether or not he is the proposer, but also to the precise identity of the proposer. Hence, in order to provide each player with the same strategic possibilities, each player can, by his bid, directly influence the choice of the proposer. An intriguing feature of the mechanism is that the Shapley value is not the outcome of just one decision, but rather emerges as the cumulative outcome of both the proposal and the bid.

The mechanism is defined recursively. If there is only one player, then he just obtains the value of his stand-alone coalition. It is also useful to describe the bidding mechanism with only two players. It is a three-stage game. First, each player makes a bid to the other. The proposer is then chosen as the player making the highest bid. If the bids are equal the proposer is chosen randomly. The proposer pays the promised bid to her partner. In the second stage, the proposer makes an offer to the other player for him to join her. In the final stage, the player who is not the proposer either accepts or rejects the offer. If he accepts, the grand coalition is formed and the proposer collects the value generated by it while paying the offer to the other player. If the proposal is rejected each player is left on his own, and hence each obtains the value of the stand-alone coalition (minus or plus the bid paid previously). Once we know the rules of a two-player bidding mechanism, we can define the mechanism for three players, and so on. Assuming that we know the rules of the bidding mechanism when played by at most $n - 1$ players, we now define the game for n players.

First, each of the players makes a bid to each of the other players. To determine the identity of the proposer, we define the 'net bid' of a player as the difference between the sum of the bids he makes to the others minus the sum of the bids the others make to him. The net bid of a player tries to measure the difference between the incentives of this player to become the proposer (what he bids) and what the others are ready to pay him for each of them to become the proposer (what the others bid to him). The player with the highest net bid is chosen to be the proposer. If several players make the highest net bid, the proposer is chosen randomly among them. Once 'named' a proposer, she proceeds to pay the bids to the other players. Second, the proposer makes an offer to each of the other players to join her. Finally, each of the other players sequentially either accepts or rejects the offer.[9] The offer is accepted, and all the players join in the grand coalition, only if all of them accept the offer. In this case, the proposer obtains the value of the coalition, paying to the others the promised offers. If the offer is rejected, the proposer is on her own, obtaining the value of her stand-alone coalition (minus the bids she has already paid).[10] The rest of the players keep their bids and play the same game with $n-1$ players.

It is important to notice that the element of randomness in the determination of the proposer is inconsequential to our proofs. Our results still hold if ties in net bids are broken deterministically as would be the case if the highest indexed player were chosen as the proposer. Randomness is introduced only in order to prevent biased treatment of the participating players. We will return to this issue in the conclusion, when we will discuss possible extensions of our mechanism.

We now describe the *bidding mechanism* more formally. Suppose first that there is only one player $\{i\}$. Then, this player obtains the value of the stand-alone coalition (i.e., $v(i)$).

Suppose now that we know the rules of the bidding mechanism when played by at most $n-1$ players. The bidding game for a set of players $N = \{1,..., n\}$ proceeds as follows:

$t = 1$: Each player $i \in N$ makes bids b^i_j in R for every $j \neq i$. Hence, at this stage, a strategy for player i is a vector $(b^i_j)_{j\neq i}$ in R^{n-1}.

For each $i \in N$, we let $B^i = \sum_{j\neq i} b^i_j - \sum_{j\neq i} b^j_i$. Let $\alpha = \text{argmax}_i(B^i)$ where, in the case of a non-unique maximizer, α is randomly chosen among the maximizing indices. Once she has been chosen, player α pays b^α_i to every player $i \neq \alpha$.

$t = 2$: Player α makes an offer y^α_j in R to every player $j \neq \alpha$. Therefore, at this stage

[9] Note that the actual sequence of players is inconsequential. The fact that players respond in sequence rather than simultaneously is crucial for ruling out 'bad' equilibria. In bad equilibria, there are several players rejecting the proposal since whenever there is at least one rejection, a rejection by any other player is optimal (the proposal will be rejected independently of his decision).

[10] Our results hold for any specification of the outside value for the proposer as long as she obtains a payment less or equal to the value of her stand-alone coalition. See Section 7 in Hart and Mas-Colell (1996) for an interpretation of a situation in which the proposer would obtain zero if the offer is rejected, and for further discussion on this extension.

a strategy for player i is a vector $(y^i_j)_{j\neq i}$ in R^{n-1} that he will follow if he is chosen to be the proposer.

$t = 3$: The players other than α, sequentially, either accept or reject the offer. If a rejection is encountered, we say the offer is rejected. Otherwise, we say the offer is accepted.

If the offer is rejected, all players other than α proceed to play the bidding mechanism where the set of players is $N\backslash\{\alpha\}$ and player α obtains the value of her stand-alone coalition. On the other hand, if the offer is accepted, each player $i \neq \alpha$ receives $y^\alpha{}_i$ and player α obtains the value of the grand coalition minus the payments $\sum_{i\neq\alpha} y_i^\alpha$.

Given that the characteristic function is $v(.)$, the final payment for player α in case of rejection is $v(\alpha) - \sum_{i\neq\alpha} b_i^\alpha$. Final payments for the other players will be the sum of the bid $b^\alpha{}_i$ received and the outcome of the mechanism where the players are $N\backslash\{\alpha\}$. In case of acceptance of the proposal, final payment to any player i other than α is given by $y^\alpha{}_{i\,+}\, b^\alpha{}_i$, whereas player α obtains $v(N) - \sum_{i\neq\alpha} y_i^\alpha - \sum_{i\neq\alpha} b_i^\alpha$.

In order to analyze the outcome of the bidding mechanism, the following well-known characterization of the Shapley value will be useful. The Shapley value of a player i is the average of the marginal contribution of this player to the grand coalition and his Shapley values in the games where a player different from i has been removed. Or, more formally,

$$\phi_i(N) = \frac{1}{n}\big(v(N) - v(N\setminus\{i\})\big) + \frac{1}{n}\sum_{j\neq i}\phi_i(N\setminus\{j\}).$$

This equation has been previously used by Maschler and Owen (1989) and Hart and Mas-Colell (1989). Furthermore, note that it provides a convenient recursive definition of the Shapley value. Defining $\phi_i(\{i\}) = v(i)$ for every i, the previous equation characterizes the Shapley value for every game (N, v).

Theorem 1*. The bidding mechanism implements the Shapley value of the zero-monotonic game* (N, v) *in SPE.*

Proof. The proof proceeds by induction on the number of players n. The theorem holds for $k = 1$, since for a one-player game, the value of his stand-alone coalition is the Shapley value.

We now assume that the theorem holds for $k = n - 1$ and show that it also holds for $k = n$. We take $N = \{1,\ldots, n\}$. We first prove that the Shapley value payoff is indeed an equilibrium outcome. We explicitly construct an SPE that yields the Shapley value as an SPE outcome. Consider the following strategies:

At $t = 1$, each player i, $i \in N$, announces $b_j^i = \phi_j(N) - \phi_j(N \setminus \{i\})$, for every $j \neq i$.
At $t = 2$, player i, $i \in N$, if he is the proposer, offers $y_j^i = \phi_j(N \setminus \{i\})$ to every $j \neq i$.
At $t = 3$, player i, $i \in N$, if player $j \neq i$ is the proposer, accepts any offer greater than or equal to $\phi_i(N\setminus\{j\})$ and rejects any offer strictly smaller than $\phi_i(N\setminus\{j\})$.

It is clear that these strategies yield the Shapley value for any player who is not the proposer, since $x^\alpha{}_i = b^\alpha{}_i + y^\alpha{}_i = \phi_i(N)$, for $i \neq \alpha$. Moreover, given that following the strategies the grand coalition is formed, the proposer also obtains her Shapley value.

We now show that all net bids B^i are equal to zero. Following the above mentioned strategies,

$$B^i = \sum_{j\neq i} b_j^i - \sum_{j\neq i} b_i^j = \sum_{j\neq i}\left(\phi_j(N) - \phi_j(N \setminus \{i\})\right) - \sum_{j\neq i}\left(\phi_i(N) - \phi_i(N \setminus \{j\})\right).$$

By the balanced contributions property (see Myerson, 1980)

$$\phi_j(N) - \phi_j(N \setminus \{i\}) = \phi_i(N) - \phi_i(N \setminus \{j\})$$

and hence $B^i = 0$.

To check that the previous strategies constitute an SPE note, first, that the strategies at $t = 2$ and $t = 3$ are best responses as long as $v(N) - v(i) \geq \sum_{j\neq i} \phi_j(N \setminus \{i\}) = v(N \setminus \{i\})$. Indeed, in the case of rejection, a proposer i obtains $v(i)$ and the players $j \neq i$ play the bidding mechanism where $N\setminus\{i\}$ is the set of players; by the induction argument, the outcome of this game is the Shapley value vector $(\phi_j(N\setminus\{i\}))_{j\neq i}$. Consider now the strategies at $t = 1$. If player i increases his total bid $\sum_{j\neq i} b_j^i$, he will be chosen as the proposer with certainty, but his payoff will decrease. If he decreases his total bid another player will propose, and player i's payoff would still equal his Shapley value. Finally, any change in his bids that leaves the total bid constant will influence the identity of the proposer but will not alter player i's payoff.

We now show that any SPE yields the Shapley value. We proceed by a series of claims:

Claim (a). In any SPE, at $t = 3$, all players other than the proposer α accept the offer if $y^\alpha{}_I > \phi_i(N\setminus\{\alpha\})$ for every player $i \neq \alpha$. Moreover, if $y^\alpha{}_i < \phi_i(N\setminus\{\alpha\})$ for at least some $i \neq \alpha$, then the offer is rejected.

Note that in the case of rejection, by the induction argument the payoff to a player $i \neq \alpha$ is $\phi_i(N\setminus\{\alpha\})$. We denote the last player that has to decide whether to accept or reject the offer, at $t = 3$, by β. If the game reaches player β, i.e., there has been no previous rejection, his optimal strategy involves accepting any offer higher than $\phi_\beta(N\backslash\{\alpha\})$ and rejecting any offer lower than $\phi_\beta(N\backslash\{\alpha\})$. The second to last player (denoted by $\beta - 1$) anticipates the reaction of player β. Hence, if $y^\alpha{}_{\beta-1} > \phi_{\beta-1}(N\backslash\{\alpha\})$ and $y^\alpha{}_\beta > \phi_\beta(N\backslash\{\alpha\})$, and the game reaches player $\beta - 1$, he will accept the offer. If $y^\alpha{}_{\beta-1} < \phi_{\beta-1}(N\backslash\{\alpha\})$ and $y^\alpha{}_\beta > \phi_\beta(N\backslash\{\alpha\})$, he will reject the offer. If $y^\alpha{}_\beta < \phi_\beta(N\backslash\{\alpha\})$, player $\beta - 1$ is indifferent to accepting or rejecting any offer $y^\alpha{}_{\beta-1}$, since he knows that player β is bound to reject the offer should the game reach him. In any case, the offer is rejected. We can go backwards using the same argument to prove claim (a).

Claim (b). If $v(N) > v(N\backslash\{\alpha\}) + v(\alpha)$, the only SPE of the game that starts at $t = 2$ is the following: At $t = 2$, player α offers $y^\alpha{}_i = \phi_i(N\backslash\{\alpha\})$ to all $i \neq \alpha$; at $t = 3$, every player $i \neq \alpha$ accepts any offer $y^\alpha{}_i \geq \phi_i(N\backslash\{\alpha\})$ and rejects the offer otherwise.

If $v(N) = v(N\backslash\{\alpha\}) + v(\alpha)$ there exist SPE in addition to the previous one. Any set of strategies where, at $t = 2$, the proposer offers $y^\alpha{}_j \leq \phi_j(N\backslash\{\alpha\})$ to a particular player $j \neq \alpha$ and, at $t = 3$, the player j rejects any offer $y^\alpha{}_j \leq \phi_j(N\backslash\{\alpha\})$, also constitutes an SPE.

In all the SPE of this subgame, the final payoffs to players α and $i \neq \alpha$ are $v(N) - v(N\setminus\{\alpha\}) - \sum_{j\neq\alpha} b^\alpha_j$ and $\phi_i(N\backslash\{\alpha\}) + b^\alpha{}_i$, respectively.

It is easy to see that the proposed strategies constitute an SPE. Suppose now that $v(N) > v(N\backslash\{\alpha\}) + v(\alpha)$. In that case, rejection of the offers made by player α cannot be part of an SPE. In such a case, player α receives $v(\alpha)$. She can improve her payoff by offering $\phi_i(N\backslash\{\alpha\}) + \varepsilon/(n - 1)$ to every $i \neq \alpha$, with $\varepsilon < v(N) - v(N\backslash\{\alpha\}) - v(\alpha)$ and $\varepsilon > 0$ so that her offers are accepted (by (a)). Therefore, an SPE requires acceptance of the proposal. This implies $y^\alpha{}_i \geq \phi_i(N\backslash\{\alpha\})$ for all $i \neq \alpha$. However, an offer such that $y^\alpha{}_j > \phi_j(N\backslash\{\alpha\})$ for some $j \neq \alpha$ cannot be part of an SPE, since α could still offer $\phi_i(N\backslash\{\alpha\}) + \varepsilon/(n - 1)$ to every $i \neq \alpha$, with $\varepsilon < y^\alpha{}_j - \phi_j(N\backslash\{\alpha\})$ and $\varepsilon > 0$. These offers are accepted and α's payoff increases. Hence, $y^\alpha{}_i = \phi_i(N\backslash\{\alpha\})$ for all $i \neq \alpha$ at any SPE. Finally, acceptance of the proposals implies that, at $t = 3$, every agent $i \neq \alpha$ accepts an offer if $y^\alpha{}_i \geq \phi_i(N\backslash\{\alpha\})$.

If $v(N) = v(N\backslash\{\alpha\}) + v(\alpha)$, then the proposer has to offer at least $\sum_{j\neq\alpha} \phi_j(N\setminus\{\alpha\}) = v(N\backslash\{\alpha\})$ for the offer to be accepted by every other player. By the same argument as in the previous case, every equilibrium in which the offer is accepted necessarily involves a proposal of exactly $\phi_j(N\backslash\{\alpha\})$ for every $j \neq \alpha$. Given that the proposer obtains $v(\alpha)$ in case of rejection, any offer that leads to a rejection would be an SPE as well.

Notice that *following* the first strategies, the offer is accepted and the grand coalition is formed, while the second strategies imply that the proposer is left on her own. Given that the last strategies are SPE only when $v(N) = v(N\backslash\{\alpha\}) + v(\alpha)$, it is easy to check that the final payoffs are those stated in the claim.

Claim (c). In any SPE, $B^i = B^j$ for all i and j and hence $B^i = 0$ for all i in N.

Denote $\Omega = \{i \in N | B^i = \text{Max}_j\ (B^j)\}$. If $\Omega = N$ the claim is satisfied since $\sum_{i \in N} B^i = 0$. Otherwise, we can show that any player i in Ω can change his bids so as to decrease the sum of payments in case he wins. Furthermore, these changes can be made without altering the set Ω. Hence, he maintains the same probability of winning, and obtains a higher expected payoff. Take some player $j \notin \Omega$. Let player $i \in \Omega$ change his strategy by announcing: $b'^{i}_{k} = b^{i}_{k} + \delta$ for all $k \in \Omega$ and $k \neq i$; $b'^{i}_{j} = b^{i}_{j} - |\Omega|\, \delta$; and $b'^{i}_{l} = b^{i}_{l}$ for all $l \notin \Omega$ and $l \neq j$. The new net bids are: $B'^{i} = B^{i} - \delta$; $B'^{k} = B^{k} - \delta$ for all $k \in \Omega$ and $k \neq i$; $B'^{j} = B^{j} + |\Omega|\delta$ and $B'^{l} = B^{l}$ for all $l \notin \Omega$ and $l \neq j$. If δ is small enough, so that $B^{j} + |\Omega|\delta < B^{i} - \delta$ (remember that $B^{j} < B^{i}$), then $B'^{l} < B'^{i} = B'^{k}$ for all $l \notin \Omega$ (including j) and for all $k \in \Omega$. Therefore, Ω does not change. However, $\sum_{h \neq i} b^i_h - \delta < \sum_{h \neq i} b^i_h$.

Claim (d). In any SPE, each player's payoff is the same regardless of who is chosen as the proposer.

We already know that all the bids B^i are the same. If player i would strictly prefer to be the proposer, he could improve his payoff by slightly increasing one of his bids b^i_j. Similarly, if player i would strictly prefer that some other player j were the proposer, he could improve his payoff by decreasing b^i_j. The fact that player i does not do so in equilibrium means that he is indifferent to the proposer's identity.

Claim (e). In any SPE, the final payment received by each of the players coincides with his Shapley value.

Note first that, if player i is the proposer, his final payoff is given by: $x^i_i = v(N) - v(N \setminus \{i\}) - \sum_{j \neq i} b^i_j$. On the other hand, if player $j \neq i$ is the proposer, the final payoff of player i is given by: $x^j_i = \phi_i(N \setminus \{j\}) + b^j_i$. Therefore, the sum of payoffs to player i over all possible choices of the proposer is given by:

$$\sum_j x^j_i = \left(v(N) - v(N \setminus \{i\}) - \sum_{j \neq i} b^i_j \right) + \sum_{j \neq i} \left(\phi_i(N \setminus \{j\}) + b^j_i \right) =$$

$$v(N) - v(N \setminus \{i\} + \sum_{j \neq i} \phi_i(N \setminus \{j\}) - B^i = v(N) - v(N \setminus \{i\}) + \sum_{j \neq i} \phi_i(N \setminus \{j\}) = n\phi_i(N),$$

Moreover, since player i is indifferent to all possible choices of the proposer, we have $x^j_i = x^k_i$ for all j, k. Therefore $x^j_i = \phi_i(N)$ for all j in N. ■

The theorem, in addition to showing that the mechanism indeed realizes the Shapley value, provides us with the explicit form of the equilibrium strategies. The ease by which these strategies can be computed adds further credibility to our results and helps in the actual implementation of the mechanism. At equilibrium, the bid of player i to player j is $\phi_j(N)-\phi_j(N\setminus\{i\})$. The balanced contributions property (see Myerson, 1980) implies that the bid can also be expressed as $\phi_i(N)-\phi_i(N\setminus\{j\})$, which is the contribution of player j to the Shapley value of player i. In particular, the bids are symmetric: player i bids for j just as much as player j bids for i. Furthermore, the determination of the offers is also simple. If player i is the proposer, he offers $\phi_j(N\setminus\{i\})$ to any other player j. The offer reflects the outside options of the players other than the proposer. Due to the recursive nature of our mechanism, these options are given by their Shapley value in the game without the proposer. Finally, notice that if the game is strictly zero-monotonic[11] not only is the equilibrium outcome unique, but the equilibrium strategies are unique as well. This eliminates problems of coordination among the players.

As we pointed out in the informal description of the mechanism, Theorem 1 holds if proposer α obtains a payment $u(\alpha)$ lower than $v(\alpha)$ in case her offer is rejected. This is a more reasonable assumption in those circumstances in which the technology is not replicable. In such a case $v(S)$ represents the payoff to the partners in S only if they have access to the technology. If $u(i) < v(i)$ for every i in N, then the equilibrium strategies are unique even if the game is zero-monotonic and not strictly zero-monotonic.

A further advantage of the mechanism is that it is finite in contrast to the infinite horizon mechanisms that implement the Shapley value in stationary SPE. Moreover, at the proposed equilibrium strategies, only the first stage of the game is played, with the proposal made by the chosen proposer accepted by the other players.

We can modify our mechanism by replacing the bidding stage with a random selection of the proposer. Once the proposer is determined, the game proceeds similarly to our mechanism with the only difference being that in case of rejection the new proposer is randomly selected from the remaining players. This modified mechanism coincides with the Mas-Colell (1988) and Hart and Mas-Colell (1996) (with removal probability equal to one) construction. In this mechanism, however, the equilibrium payoffs yield the Shapley value in expected terms only.

4. Implementation of the Weighted Shapley Values

The weighted Shapley value emerges out of considering non-symmetric divisions of the surplus. It is defined in Shapley (1953) by stipulating an exogenously given system of weights $w \in R^n_{++}$. Each unanimity game is assigned a value by having

[11] We say that a game is strictly zero-monotonic if $v(S) + v(\{i\}) < v(S \cup \{i\})$ for any subsets $S \subseteq N$ with $i \notin S$ and $S \neq \Phi$.

agent i receive the share $w^i / \sum_{j\in N} w^j$ of the unit. The *w-weighted Shapley value* of a game is defined as the linear extension of this operator to the game. We denote by $\phi_{wi}(N)$ the w-weighted Shapley value of player i in the cooperative game (N, v).

A convenient way to express the weighted Shapley value is through the weighted potential function $P_w(N)$ defined in Hart and Mas-Colell (1989).[12] The w-weighted potential $P_w(N)$ is the unique function from the set of games into R that satisfies $P_w(\Phi) = 0$ and $\sum_{i\in N} w^i D^i P_w(N) = v(N)$, where $D^iP_w(N) = P_w(N) - P_w(N\backslash\{i\})$. This function satisfies: $w^iD^iP_w(N) = \phi_{wi}(N)$. Furthermore,

$$P_w(N) = \frac{1}{\sum_{j\in N} w^j}\left[v(N) + \sum_{j\in N} w^j P_w(N\setminus\{j\})\right].$$

The weighted Shapley value, as the Shapley value, can be calculated using a recursive procedure. The role played by this formula in the proof of Theorem 2 is similar to the role played by the recursive formula characterizing the Shapley value in the proof of Theorem 1:

***Lemma 1**. The weighted Shapley value of player i satisfies the equality*:

$$\phi_{wi}(N) = \frac{1}{\sum_{j\in N} w^j}\left[w^i(v(N) - v(N\setminus\{i\})) + \sum_{j\neq i} w^j \ \phi_{wi}(N\setminus\{j\})\right].$$

Proof. The weighted Shapley value of player i satisfies:

$$\begin{aligned}
\phi_{wi}(N) &= w^i\left[P_w(N) - P_w(N\setminus\{i\})\right] \\
&= w^i \frac{1}{\sum_{j\in N} w^j}\left[v(N) + \sum_{j\in N} w^j P_w(N\setminus\{j\}) - \sum_{j\in N} w^j P_w(N\setminus\{i\})\right] = \\
&= \frac{1}{\sum_{j\in N} w^j}\Bigg[w^i v(N) + \sum_{j\neq i} w^j \Big(w^i P_w(N\setminus\{j\}) - w^i P_w(N\setminus\{i,j\}) \\
&\qquad - w^i P_w(N\setminus\{i\}) + w^i P_w(N\setminus\{i,j\})\Big)\Bigg] \\
&= \frac{1}{\sum_{j\in N} w^j}\left[w^i v(N) + \sum_{j\neq i} w^j \ \phi_{wi}(N\setminus\{j\}) - w^i \sum_{j\neq i} \ \phi_{wj}(N\setminus\{i\})\right] \\
&= \frac{1}{\sum_{j\in N} w^j}\left[w^i(v(N) - v(N\setminus\{i\})) + \sum_{j\neq i} w^j \ \phi_{wi}(N\setminus\{j\})\right]. \qquad \blacksquare
\end{aligned}$$

We will now indicate how to modify our original bidding mechanism in order to

[12] Again, we omit the constant v and write for short $\phi_{wi}(N)$ or $P_w(N)$ instead of $\phi_{wi}(N, v)$ or $P_w(N, v)$.

obtain as an equilibrium outcome any weighted Shapley value. The only difference is in the construction of the weighted net bids B_w^i. The determination of net bids incorporates the vector of weights $w \in R^n{}_{++}$ by having $B_w{}^i = \sum_{j \neq i} w^i b_j^i - \sum_{j \neq i} w^j b_i^j$. Other than that change, the *weighted bidding mechanism* proceeds like the bidding mechanism. Intuitively we weigh each bid differently, according to the exogenously given weight of the person making the bid.

Theorem 2. *The weighted bidding mechanism implements the weighted Shapley value of the zero-monotonic game (N, v) in SPE.*

The proof of Theorem 2 is similar to the proof of Theorem 1.

Finally, note that we can implement the weighted Shapley value in expected terms by using a simpler mechanism (similar to the Mas-Colell (1988) and Hart and Mas-Colell (1996), construction for the Shapley value). Given a system of weights $w \in R^n{}_{++}$, we replace the bidding stage by a random choice of the proposer, where the probability of player i to be chosen as the proposer equals $w^i / \sum_{j \in N} w^j$ (rather than $1/n$).

5. General Transferable Utility Games and the Formation of Coalitions

The only requirement we have imposed so far on the cooperative environment is that of zero-monotonicity. Zero-monotonic environments might still violate super-additivity. Therefore the (weighted) bidding mechanism implements the (weighted) Shapley value even in some non super-additive settings. This result however is not entirely satisfactory since the outcome while coinciding with the Shapley value might not be ‘really’ efficient. The sum of payments would indeed equal $v(N)$, yet $v(N)$ might not be the maximal payoff the players could obtain. Note that in non super-additive environments it might be possible for the players to obtain a sum of payments that exceeds $v(N)$ by splitting up into two or more coalitions.

One way to resolve this difficulty might be to consider the super-additive cover of the environment. If we apply our mechanism to the super-additive cover of the environment rather than to the original environment, the equilibria outcomes would coincide with the Shapley value of the super-additive cover. A possible disadvantage of this approach is that a player (the proposer) is able to collect rents from a coalition of which she is not an active member. In other words, a player can act as a ‘principal’ for a coalition formed by other players.[13]

One way to avoid the use of ‘principals’ is to modify our mechanism. The new generalized bidding mechanism would generate a coalition structure in which proposers would receive (when there is no rejection) just the value of the coalition

[13] See Pérez-Castrillo (1994) and Pérez-Castrillo and Wettstein (2000) for the use of principals to realize cooperative outcomes.

to which they belong. In this mechanism the players would not only share the surplus but would also form coalitions in a sequential way. We show that at any SPE, the coalitions formed will constitute an efficient coalition structure and the final payments of the players will coincide with the Shapley value of the super-additive cover of the environment.

Before proceeding with the formal description of the generalized bidding mechanism we introduce the following notation. The *super-additive* (SA) *cover* of a cooperative game in characteristic form (N, v), is denoted by (N, V). The value $V(S)$, for $S \subseteq N$, is defined by: $V(S) = \mathrm{Max}_{\pi}\{\sum_{S\in\pi} v(S) \mid \pi \text{ is a partition of } S\}$.

We denote the Shapley value of player i in the SA cover of (N, v) by $\Theta_i(N)$, and similarly for the values $\Theta_i(S)$ of subsets S of N.

We know that:

$$\Theta_i(N) = \frac{1}{n}\big(V(N) - V(N \setminus \{i\})\big) + \frac{1}{n}\sum_{j\neq i}\Theta_i(N \setminus \{j\}).$$

A partition π such that $V(N) = \sum_{S\in\pi} v(S)$ is called an *efficient partition* for N.

The generalized bidding mechanism (GBM) is similar to the bidding mechanism. The only difference is that in the GBM, the proposer, in addition to offering a vector of payments to all the other players, also chooses a coalition she wants to form and be a member of. Hence, an offer by the proposer consists of a payments vector and a coalition. The offer is accepted if all the other players agree. In case of acceptance the coalition is formed, the proposer collects the value of that coalition and the players outside the coalition proceed to play the same game again among themselves. In the case of rejection all the players other than the proposer play the same game again.

Formally, if there is only one player $\{i\}$, she obtains the value of the stand-alone coalition. Given the rules of the game when played by at most $n - 1$ players, the game for $N = \{1,..., n\}$ players proceeds as follows:

$t = 1$: Each player $i \in N$ makes bids b^i_j in R for every $j \neq i$.

Player α is chosen as in the bidding mechanism. She pays b^{α}_{i} to every player $i \neq \alpha$.

$t = 2$: Player α chooses a coalition S_α with $\alpha \in S_\alpha$ and makes an offer y^{α}_{i} in R to every player $i \neq \alpha$.
$t = 3$: The players other than α, sequentially, either accept or reject the offer. If an agent rejects it, then the offer is rejected. Otherwise, the offer is accepted.

If the offer is accepted, each player $i \neq \alpha$ receives y^{α}_{i} and player α receives the value of the coalition S_α minus the payments $\sum_{i\neq\alpha} y^{\alpha}_{i}$. After this, players in $N \setminus S_\alpha$

proceed to play the GBM again among themselves. (Therefore, final payment to a player $i \in S_\alpha \backslash \{\alpha\}$ is $y^{\alpha}_{i} + b^{\alpha}_{i}$, player α receives $v(S_\alpha) - \sum_{i \neq \alpha} y_i^\alpha - \sum_{i \neq \alpha} b_i^\alpha$, and the final payment for a player $i \in N \backslash S_\alpha$ will be the sum of the bid b^{α}_{i}, the offer y^{α}_{i}, and the outcome of the GBM where the players are $N \backslash S_\alpha$.) On the other hand, if the offer is rejected, all players other than α proceed to play the GBM where the set of players is $N \backslash \{\alpha\}$ and player α receives the value of her stand-alone coalition.

Theorem 3. *The generalized bidding mechanism implements the Shapley value of the SA cover of the game (N, v).*

Proof. The arguments, in part, are very similar to those used in Theorem 1, thus we emphasize just the new features of this proof and otherwise rely on the reasoning employed in Theorem 1.

It is easy to see that the theorem holds for $k = 1$. We assume that it holds for $k = n - 1$ and then consider the following strategies:

At $t = 1$, each player i, $i \in N$, announces $b^i_j = \Theta_j(N) - \Theta_j(N \backslash \{i\})$, for every $j \neq i$.

At $t = 2$, player i, $i \in N$, if she is the proposer, chooses a coalition S_i such that $S_i \in \text{Argmax}_{S \subseteq N} \{v(S) + V(N \backslash S) \mid i \text{ in } S\}$ and offers $y^i_j = \Theta_j(N \backslash \{i\})$ to every $j \in S_i \backslash \{i\}$ and $y^i_j = \Theta_j(N \backslash \{i\}) - \Theta_j(N \backslash S_i)$ to every $j \notin S_i$.

At $t = 3$, player i, $i \in N$, if player $j \neq i$ is the proposer and $i \in S_j$, accepts any offer greater than or equal to $\Theta_i(N \backslash \{j\})$ and rejects it otherwise. If player $j \neq i$ is the proposer and $i \notin S_j$, player i accepts any offer greater than or equal to $\Theta_i(N \backslash \{j\}) - \Theta_i(N \backslash S_j)$ and rejects it otherwise.

Following these strategies, the proposer selects a coalition S_α that is part of an efficient partition.[14] Also, the induction argument ensures that, in the game that follows among the players in $N \backslash S_\alpha$, player $i \notin S_\alpha$ will obtain $\Theta_i(N \backslash S_\alpha)$. It is then easy to see that the previous strategies yield $\Theta_i(N)$ to any player i.

To prove that the previous strategies constitute an SPE, note, first, that the strategy at $t = 3$ is a best response for any player different from the proposer by the same argument we used in Theorem 1. At $t = 2$, given the rejection criteria used by the other players, if player i is the proposer, she chooses a subset S_i that maximizes:

$$v(S_i) - \sum_{j \in S_i \backslash \{i\}} \Theta_j(N \backslash \{i\}) - \sum_{j \notin S_i} \left\lfloor \Theta_j(N \backslash \{i\}) - \Theta_j(N \backslash S_i) \right\rfloor = v(S_i) + V(N \backslash S_i) - \mathrm{V}(N \backslash \{i\}).$$

Therefore, the proposed strategy is optimal. Finally, an argument similar to

[14] It can be easily shown that $\mathrm{V}(N) = \text{Max}_{S \subseteq N} \{v(S) + V(N \backslash S) \mid \alpha \text{ in } S\}$, for any player α in N, hence when the proposer chooses the best possible coalition to be a member of, she is choosing a coalition that forms part of an efficient partition.

the one in the proof of Theorem 1 demonstrates the optimality of the strategies at $t = 1$.

To show that any SPE yields the Shapley value, we proceed by a series of claims. We state the claims without proof, since they are similar to those in Theorem 1. To simplify notation, we denote the 'effective offer' to player $i \neq \alpha$ in stage 2, when player α is the proposer by z^{α}_{i}, and define it as $z^{\alpha}_{i} = y^{\alpha}_{i}$ if $i \in S_{\alpha} \backslash \{\alpha\}$ and $z^{\alpha}_{i} = y^{\alpha}_{i} + \Theta_{i}(N \backslash S_{\alpha})$ if $i \notin S_{\alpha}$. By the induction argument, the effective offer is the total payment (without taking into account the bid already received) that a player will receive (at equilibrium) if the offer is accepted.

Claim (a). In any SPE, at $t = 3$, any player $j \neq \alpha$ accepts the offer if z^{α}_{j} is strictly greater than $\Theta_{i}(N \backslash \{\alpha\})$ for every player $i \neq \alpha$. Moreover, if $z^{\alpha}_{i} < \Theta_{i}(N \backslash \{\alpha\})$ for at least some $i \neq \alpha$, then the offer is rejected.

Claim (b). If the coalition $\{\alpha\}$ is not part of any efficient partition, then in any SPE of the game that starts at $t = 2$, α will choose a coalition S_{α} that is part of an efficient partition. Furthermore, player α will announce offers such that $z^{\alpha}_{i} = \Theta_{i}(N \backslash \{\alpha\})$ for any player $i \neq \alpha$. Finally, at $t = 3$, every player $i \neq \alpha$ accepts any offer such that $z^{\alpha}_{i} \geq \Theta_{i}(N \backslash \{\alpha\})$.

If the coalition $\{\alpha\}$ is part of any efficient partition, there exist other equilibria in addition to the previous ones. Any set of strategies where, at $t = 2$, the proposer makes offers such that $z^{\alpha}_{j} \leq \Theta_{j}(N \backslash \{\alpha\})$ to a particular player $j \neq \alpha$ and, at $t = 3$, the player j rejects any effective offer less than or equal to $\Theta_{j}(N \backslash \{\alpha\})$, also constitute an SPE.

In all the SPE of this subgame, the payoffs (taking into account the continuation of the game after S_{α} has been formed) to players α and $i \neq \alpha$ are $V(N) - V(N \backslash \{\alpha\}) - \sum_{j \neq \alpha} b^{\alpha}_{j}$ and $\Theta_{i}(N \backslash \{\alpha\}) + b^{\alpha}_{i}$, respectively.

(Notice that following both types of strategies an efficient partition is formed.)

Claim (c). In any SPE, $B^{i} = 0$ for all i in N.

Claim (d). In any SPE, each player's payoff is the same regardless of who is chosen as the proposer.

Claim (e). In any SPE, the final payment received by each of the players coincides with his Shapley value in the SA cover. ■

Theorem 3 shows that when facing environments where forming the grand coalition might not be efficient, it is possible to employ a generalized version of our original bidding mechanism that allows both that an efficient partition can be formed and that the surplus can be shared in a 'reasonable' way. If the game is super-additive, the generalized version yields the same outcome as the bidding mechanism. It is however important to notice that, if the game is not super-additive but the grand coalition is efficient, this coalition is formed under both mechanisms although the sharing of the surplus will be different. The reason is that the Shapley value of the super-additive cover usually does not coincide with the Shapley value of the game if the game is not super-additive.

Our GBM provides support for the use of the Shapley value of the SA cover as the generalization of the Shapley value for games in which it is efficient to form coalition structures which are different from the grand coalition. The GBM implements the Shapley value of the SA cover by simultaneously providing a *bidding and coalition formation game*. To the best of our knowledge, this is the first paper that supports this solution concept. Aumann and Dréze (1974) study games with a (*given*) coalition structure and define a value that assigns to each player his Shapley value in the coalition he belongs to. Under this concept, the payoff to any player does not depend upon his contribution to coalitions other than his coalition. The Shapley value of the super-additive cover takes into account not only the contribution of a player to the coalition he belongs to in an efficient structure, but also his potential contribution to any other coalition.[15]

6. Conclusion

The object of this paper was to construct a simple non-cooperative mechanism to realize a sharing of the surplus in a cooperative environment. The mechanism we use basically consists of two distinct stages of play: a bidding stage, at the end of which a winner is determined, followed by a proposal stage where the winner offers a sharing of the surplus. In the case where the proposal is rejected, the same game is played again by the players except for the proposer. We show that the payoff outcome of the subgame perfect equilibria of this game always coincides with the Shapley value of the game. Moreover, the strategies played by the players at equilibrium are simple and natural. We also showed that a natural modification of the mechanism implements the weighted Shapley value. Finally, we have introduced a simple generalization of the bidding mechanism that handles situations where the grand coalition might not be efficient. By playing the game, the players form, at equilibrium, an efficient coalition structure and share the surplus according to the Shapley value of the super-additive cover of the environment.

[15] Owen (1977) and Hart and Kurz (1983) also propose a coalition structure value to every game and every coalition structure. However, in their approach, the coalition structure serves only as a bargaining tool to increase the payoff of the members of the coalitions. At the end, all the players join the grand coalition.

These mechanisms provide strong support for applying the Shapley value to situations where cooperation is needed to obtain an efficient outcome. It might be also used for a variety of cost allocation, revenue sharing, or partnership dissolution problems.

The general approach taken in this paper may yield ways to provide non-cooperative foundations for other cooperative solution concepts for transferable utility games or for cost-sharing methods. However, the extension of our approach to non-transferable utility games is problematic. There exist several extensions of the Shapley value to non-transferable utility games proposed by Harsanyi (1963), Shapley (1969), and Maschler and Owen (1989, 1992). Dagan and Serrano (1998) have shown that randomness is a necessary component in a mechanism designed to implement any of these extensions. Since the element of randomness in our mechanism (i.e., the tie-breaking rule) is inconsequential to proving our results, it seems that the approach taken in this paper would fail to implement the existing extensions of the Shapley value.

References

Aumann, R. and J. Dréze (1974), 'Cooperative games with coalition structure', *International Journal of Game Theory* **3**, 217–237.

Crawford, V.P. (1979), 'A procedure for generating Pareto-efficient egalitarian-equivalent allocations', *Econometrica* **47**, 49–60.

Dagan, N. and R. Serrano (1998), 'Invariance and randomness in the Nash program for coalitional games', *Economics Letters* **58**, 43–49.

Dasgupta, A. and Y. S. Chiu (1996), 'On implementation via demand commitment games', *International Journal of Game Theory* **27**, 161–189.

Evans, R.A. (1996), 'Value, consistency, and random coalition formation', *Games and Economic Behavior* **12**, 68–80, doi:10.1006/game.1966.0005.

Evans, R.A. (1997), 'Coalitional bargaining with competition to make offers', *Games and Economic Behavior* **19**, 211–220, doi:10.1006/game.1997.0553.

Gul, F. (1989), 'Bargaining foundations of Shapley value', *Econometrica* **57**, 81–95.

Gul, F. (1999), 'Efficiency and immediate agreement: A reply to Hart and Levy', *Econometrica* **67**, 913–917.

Harsanyi, J.C. (1963), 'A simplified bargaining model for the *n*-person cooperative game', *International Economic Review* **4**, 194–220.

Hart, S. and M. Kurz (1983), 'Endogenous formation of coalitions', *Econometrica* **51**, 1047–1064.

Hart, S. and Z. Levy (1999), 'Efficiency does not imply immediate agreement', *Econometrica* **67**, 909–912.

Hart, S. and A. Mas-Colell (1989), 'Potential, value, and consistency', *Econometrica* **57**, 589–614.

Hart, S. and A. Mas-Colell (1996), 'Bargaining and value', *Econometrica* **64**, 357–380.

Hart, O. and J. Moore (1990), 'Property rights and the nature of the firm', *Journal of Political Economy* **98**, 1119–1158.

Krishna, V. and R. Serrano (1995), 'Perfect equilibria of a model of N-person non-cooperative bargaining', *International Journal of Game Theory* **24**, 259–272.

Maschler, M. and G. Owen (1989), 'The consistent Shapley value for hyperplane games', *International Journal of Game Theory* **18**, 389–407.

Maschler, M. and G. Owen (1992), 'The consistent Shapley value for games without side payments', in R. Selten (ed.), *Rational Interaction*, New York: Springer-Verlag, pp. 5–12.

Mas-Colell, A. (1988), 'Algunos comentarios sobre la teoría cooperativa de los juegos', *Cuadernos Económicos de ICE* **40**, 143–161.

Myerson, R.B. (1980), 'Conference structures and fair allocation rules', *International Journal of Game Theory* **9**, 169–182.

Owen, G. (1977), 'Values of games with a priori unions', in R. Hein and O. Moeschlin (eds.), *Essays in Mathematical Economics and Game Theory*, New York: Springer-Verlag, pp. 76–88.

Pérez-Castrillo, D. (1994), 'Cooperative outcomes through non-cooperative games', *Games and Economic Behavior* **7**, 428–440, doi:10:1006/game.1994.1060.

Pérez-Castrillo, D. and D. Wettstein (2000), 'Implementation of bargaining sets via simple mechanisms', *Games and Economic Behavior* **31**, 106–120, doi:10.1006/game.1999.0730.

Roth, A.E. (1988), 'Introduction to the Shapley value', in A.E. Roth (ed.), *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, Cambridge: Cambridge University Press, pp. 1–27.

Shapley, L.S. (1953), 'A value for n-person games', in H.W. Kuhn and A.W. Tucker (eds.), *Contributions to the Theory of Games II* (Ann. Math. Studies, 28), Princeton: Princeton University Press, pp. 307–317.

Shapley, L.S. (1953), 'Additive and non-additive set functions', Ph.D. thesis, Princeton: Princeton University.

Shapley, L.S. (1969), 'Utility comparison and the theory of games', in Editions du CNRS, *La Décision*, Paris, pp. 251–263

Thomson, W. (1988), 'The manipulability of the Shapley value', *International Journal of Game Theory* **17**, 101–127.

Winter, E. (1994), 'The demand commitment bargaining and snowballing cooperation', *Economic Theory* **4**, 255–273.

Free Trade Networks

Taiji Furusawa and Hideo Konishi

Paper presented at the CTN Workshop 2003. Reprinted from *Journal of International Economics* **72**(2), July 2007, 310–335, with permission from Elsevier.

The paper examines the formation of free trade agreements (FTAs) as a network formation game. We consider an n-country model in which (possibly asymmetric) countries trade differentiated industrial commodities. We show that if all countries are symmetric, the complete FTA network is pairwise stable and it is the unique stable network if industrial commodities are not highly substitutable. We also compare FTAs and customs unions (CUs) as to which of these two regimes facilitates global trade liberalization, noticing that unlike CUs, each signatory of an FTA can have another FTA without consent of other member countries.

1. Introduction

The network of preferential trade agreements (PTAs) covers most countries in a complex way. The tendency towards 'regionalism,' a movement to form regional

Furusawa gratefully acknowledges financial supports from the Fulbright Foundation, Japan Economic Research Foundation, Seimeikai Foundation, and Ministry of Education, Culture, Sports, Science and Technology of Japan (the 21st Century Center of Excellence Project on the Normative Evaluation and Social Choice of Contemporary Economic Systems). Konishi gratefully acknowledges financial support from the Japan Society for Promotion of Science research. Part of this research was conducted when Furusawa was visiting Boston University, and when Konishi was visiting RIEB at Kobe University. We thank these institutions for their hospitality and support. We are also grateful to Jim Anderson, Eric Bond, Alan Deardorff, Jonathan Eaton, Makoto Ikema, Jota Ishikawa, Ron Jones, Kazuharu Kiyono, Istv´an K´onya, Carsten Kowalczyk, Kala Krishna, Kamal Saggi, Robert Staiger, Masatoshi Tsumagari, Makoto Yano, and two anonymous referees for helpful comments. We also benefited from discussions in seminars at Boston College, Hitotsubashi University, Keio University, Kobe University, Kyoto University, La Trobe University, Osaka University, Osaka City University, Southern Methodist University, Stockholm University, Tulane University, University of New South Wales and Vanderbilt University, and the conferences which are the 2002 RIEB International Economics Conference, the 2002 European Trade Study Group meeting, the Fall 2002 Midwest International Economics meeting, and the 2002 Japanese Economic Association meeting.

trade agreements, has been steadily growing especially since 1980s (Bhagwati, 1993). Since the Treaty of Rome established the European Economic Community (EEC) in 1957, the European Union (EU) has been growing with the accession of new members. The North American Free Trade Agreement (NAFTA) has started negotiations with Latin American countries to form the Free Trade Area of the Americas. Japan has recently signed free trade agreements (FTAs) with Singapore and Mexico. The website of the World Trade Organization (WTO) on regionalism provides us with an excellent introduction to this topic.

> The vast majority of WTO members are party to one or more regional trade agreements. The surge in RTAs has continued unabated since the early 1990s. Some 250 RTAs have been notified to the GATT/WTO up to December 2002, of which 130 were notified after January 1995. Over 170 RTAs are currently in force; an additional 70 are estimated to be operational although not yet notified. By the end of 2005, if RTAs reportedly planned or already under negotiation are concluded, the total number of RTAs in force might well approach 300.
> (http://www.wto.org/english/tratop/_e/region_e/region_e.htm, August 23, 2005)
> One of the most frequently asked questions is whether these regional groups help or hinder the WTO's multilateral trading system. A committee is keeping an eye on developments.
> (http://www.wto.org/english/thewto_e/whatis_e/tif_e/bey1_e.htm, August 23, 2005)

Whether PTAs serve as 'building blocks' or 'stumbling blocks' is a central question in this topic (Bhagwati, 1993). Of course, multilateral trade liberalization efforts and PTA formation interact with each other.[1] However, putting this feature aside for a while, another important question remains. Will successive PTA formation alone effectively achieve global free trade, or will the process stop prematurely so that the world is divided into several, mutually exclusive trading blocs? If PTA formation continues until the complete FTA network is achieved, we may conclude that PTAs are 'building blocks.' But otherwise, PTAs can be 'stumbling blocks.'[2]

Ohyama (1972) and Kemp and Wan (1976) demonstrate a positive result for this 'dynamic' path problem. The so-called Kemp-Wan theorem states that member countries can appropriately adjust external tariffs and make internal transfers so that a newly formed customs union (CU) is Pareto-improving, no only

[1] Levy (1997), Krishna (1998), and Ornelas (2005c) show in their political economy models that PTA formation can hinder multilateral trade liberalization. Freund (2000b) demonstrates that countries have more incentive to form PTAs as multilateral trade negotiations lower tariffs imposed by every country. See also Bagwell and Staiger (1997a,b), Bond et al. (2001), and Ethier (1998).

[2] Bhagwati and Panagariya (1996) raise this 'PTA time-path' question. The complete FTA network may still be different from global free trade attained through multilateral trade negotiations, as Freund (2000a) demonstrates in a model where firms incur distribution network costs, for example. The complete FTA network may be more complex and inefficient ('spaghetti bowl' phenomenon) than global free trade attained through multilateral trade negotiations, as Bhagwati and Panagariya (1996) claim.

to members themselves but also to all countries in the world.[3] Successive application of this Kemp-Wan process implies that the CU expansion continues until all countries in the world are covered.[4] Although the theorem looks promising, it should be taken as an existence theorem (of a Pareto-improving CU expansion). In reality, it is extraordinarily difficult to adjust external tariffs such that each nonmember country's welfare is not reduced by the CU formation. Indeed, as Viner (1950) taught us, adverse trade-diversion effects often prevent PTAs from being Pareto improving.[5] It is far from obvious that in reality, countries always have incentives to form PTAs so that we will eventually observe the complete free trade network (global free trade). Indeed, Yi (1996) shows that even if countries are symmetric, the world would be divided into two CUs of asymmetric size when the number of countries is a realistic number.

CUs are not the only form of PTA. A PTA can take a form of FTA, such as the NAFTA, in which member countries choose their individual external tariffs without consent of other member countries unlike in the case of CU where all member countries adopt the same external tariff schedule. An important consequence of this difference, which seems to be overlooked more or less in the literature, is that under an FTA, each member country (or a subset of member countries) can sign another FTA with outside countries without consent of other member countries. Whereas in the case of CUs, such as the EU, all member countries should be involved when an outside country forms a PTA with a member country of a CU. Thus, FTAs are more flexible than CUs: A hub-and-spoke system, for example, will not appear if only CUs are allowed as PTAs.[6] In practice, CUs and FTAs co-exist in a complex manner. The hub-and-spoke system is prevalent in the world. Mexico, which is a member of the NAFTA, has FTAs with the EU, Japan, and others. The traditional approach by coalition formation games such as Yi (1996, 2000) is not rich enough to capture this feature of the world PTA configuration. Coalition formation games cannot properly address the issues of the web of FTAs, nor can they analyze the situation where CUs and FTAs co-exist.

The network formation game developed by Jackson and Wolinsky (1996) provides an appropriate framework to analyze such complex formation of PTAs. The network formation game is suitable for the analysis of FTAs. We can predict whether or not an arbitrary FTA configuration is stable. As we show in Section 4, the situation in which CUs and FTAs co-exist can also be analyzed within the same

[3] See Panagariya and Krishna (2002) for an FTA version of the Kemp-Wan theorem.

[4] Baldwin (1995) demonstrates that as a regional trading bloc expands, outside countries have more incentive to join the bloc.

[5] Krugman (1991) claims that if a 'natural' trading bloc, within which a large share of trade takes place even in the absence of a PTA, is formed, the gains from trade creation are likely to outweigh the losses from trade diversion.

[6] Kowalczyk and Wonnacott (1992) discuss the hub-and-spoke system in the argument about NAFTA. Mukunoki and Tachi (2006) investigate dynamic formation of bilateral FTAs in a three-country model.

framework. In this paper, given any FTA configuration in the world, we examine whether or not a pair of countries has an incentive to sign an FTA, and whether or not a country has an incentive to cut an existing FTA. A network that is immune to such deviations is called *(pairwise) stable network* (Jackson and Wolinsky, 1996). Then we ask if the complete FTA network is stable, and if it is, we further ask if it is unique. If the complete FTA network is a unique stable network, the world is likely to attain global free trade, building many bilateral FTAs.[7] If the complete FTA network is not stable, on the other hand, FTA formation would stop prematurely. Investigating countries' incentives to sign FTAs and deriving conditions under which the complete FTA network is stable, we hope to gain an insight into how far the worldwide movement toward FTAs continues.[8]

First, we analyze each country's incentive to sign or abandon an FTA. As Krugman (1991) and Grossman and Helpman (1995) suggest, the asymmetry of countries is an important factor when we assess countries' incentives for FTAs. Viner (1950), on the other hand, suggests that substitutability of commodities traded internationally is also an important factor. The model of this paper is general enough to allow us to observe how these factors affect a country's decision to sign an FTA. We consider the model in which the world consists of n countries that trade a numeraire good and a continuum of non-numeraire, differentiated, industrial commodities. Consumers in all countries share a common quasi-linear utility function, in which substitutability of industrial commodities is parameterized. Countries may be different in the market size (population size) and the size of the industrial good industry (measure of firms). Each of the differentiated industrial commodities is produced by one firm that belongs to one of n countries. An FTA between countries i and j simply means that countries i and j simultaneously eliminate tariffs on industrial commodities imported from each other.

Furusawa and Konishi (2004) show that when consumers have quasi-linear utility functions and all countries share the same constant-returns-to-scale

[7] To derive a definite prediction regarding the time-path to global free trade, we may need to build a dynamic network formation model with farsightedness. Mukunoki and Tachi (2006) show in a dynamic, symmetric, three-country model that under certain parameter values, only one bilateral FTA is signed in equilibrium so that global free trade is not attained. As Kennan and Riezman (1990) suggest, countries in a bilateral FTA may in some cases prefer the current situation to global free trade. Then, each member country may not sign a new bilateral FTA with an outside country since it would induce an FTA between spoke countries, effectively attaining global free trade, in the future. However, extending Mukunoki and Tachi's (2006) analysis to the case of many countries is not an easy task.

[8] Driven by the same motivation, Freund (2000c) builds a model such that each country calls out the number of countries with which it wants to have FTAs, and shows that global free trade is effectively attained as a unique Nash equilibrium. However, she seems to assume implicitly that a bilateral FTA between two countries is made effective as long as one of the countries benefits from an agreement, even if the other strictly prefers not to sign the agreement. This 'open membership' rule (see also Yi, 1996) does not seem to be appealing for discussions of FTAs. If FTAs require consent from both sides, then we will run into the multiplicity problem of Nash equilibria (see footnote 16).

production technology for each commodity they commonly produce, social welfare of a country can be represented by the sum of consumers' gross utilities and trade surplus of non-numeraire goods. An FTA with another country is likely to raise the gross utility, although the second-best effect may sometimes outweigh the benefits from tariff reduction.[9] On the other hand, the impact on the (industrial) trade surplus is generally ambiguous, and is often crucial in determining whether or not an FTA is welfare improving.

The effect on a country's trade surplus of signing an FTA with another country can be further decomposed into two: one on the trade surplus between these two countries (the direct surplus effect) and the other on the trade surplus with third countries (the third country effect). The latter effect is always positive, since the country's exports to third countries are not affected by the FTA, whereas its imports from them decrease because their commodities become relatively more expensive after the FTA. Thus, the third country effect always serves to encourage countries to sign FTAs at the costs of third countries: all other countries including existing FTA partners are hurt by these new FTAs. In contrast, the sign of the direct surplus effect depends on the two countries' characteristics such as the market and industry size, and the characteristics of their current partners. Let us consider, for example, an FTA between a highly-industrialized small country and a less-industrialized large country. The FTA increases trade flows from the former to the latter disproportionately, dramatically increasing the trade surplus of the small highly-industrialized country and decreasing that of the large less-industrialized country. The direct surplus effect for the large less-industrialized country is likely to be negative, and it may outweigh the third country effect. Consequently, the large less-industrialized country is likely to oppose to sign the FTA.[10] If two countries are similar in their characteristics, however, the direct surplus effects would be small, and the countries are likely to benefit from signing an FTA due to the third country effect.

The main results of this paper are as follows. When all countries are symmetric in the market size and the industry size, we show that the complete FTA network, the network in which any pair of countries has an FTA, is pairwise stable (Proposition 1). If commodities are highly substitutable among themselves, however, there may also be other pairwise stable networks. It is because the

[9] If tariffs have been imposed on a large portion of commodities, it may not be welfare improving to get rid of tariffs for a small portion of commodities since it enlarges distortions between these commodities and the ones with high tariffs.

[10] It is interesting to note that countries in our model have a view that Krugman (1991) calls GATT-think: '(1) Exports are good, (2) Imports are bad, (3) and other things being equal, an equal increase in imports and exports is good.' Our model gives an economic reasoning to this 'enlightened mercantilism' (see Furusawa and Konishi, 2004, for details). Bagwell and Staiger (1999a) argue that GATT's principle of reciprocity, which appears to reflect the 'enlightened mercantilism,' indeed has a sound economic role of enhancing efficiency.

difference in the number of FTA partners can create a large differential in the impacts on the direct surplus, even though all countries are symmetric in the market size and industry size. We show that if predetermined external tariff rates are small or if commodities are not highly substitutable among themselves, the complete FTA network is a unique pairwise stable network (Proposition 2). If countries are asymmetric, on the other hand, the complete FTA network may not be attained. In a special case where all industrial commodities are independent from one another, a pair of countries signs an FTA if and only if their industrialization levels are close to each other (Proposition 3).[11] This proposition implies for example that developed countries and less developed countries respectively form mutually exclusive trading blocs. We also compare FTAs and CUs as to which of these two regimes facilitate global trade liberalization. We find that if all countries are symmetric, and if industrial commodities are not highly substitutable among themselves, a pair of countries has less incentive to form a new FTA if either of them is a member of a CU as opposed to an FTA (Proposition 4). If countries are asymmetric, on the other hand, the CU formation averages out member countries' industrialization levels, which may help further PTA formation. We illustrate this possibility in the case of mutually independent industrial commodities.

An independent work by Goyal and Joshi (2006) also investigates the FTA formation as a network formation game, and obtains the result that the complete FTA network is pairwise stable (the counterpart of our Proposition 1). Our model is richer in some important aspects, enabling us to obtain further insights on incentives to sign FTAs. In particular, their model assumes that firms produce a homogeneous good, whereas ours has an industry with differentiated commodities whose substitutability is parameterized. As briefly discussed above, substitutability among differentiated commodities plays an important role in determining the global FTA configuration. In addition, our model is more suitable for the analysis of asymmetric countries than theirs. The main part of their analysis assumes that all countries are symmetric with respect to the (Cournot-oligopolistic) market size and the number of domestic firms, whereas ours are more flexible so that we obtain such a result as Proposition 3 in the case of asymmetric countries. We also discuss the difference between FTAs and CUs as to which of them facilitates global trade liberalization in a higher degree.

[11] Furusawa and Konishi (2005) show that Propositions 1 and 2 in this paper can be generalized to the case of asymmetric countries if transfers between FTA signatories are allowed. With transfers, a pair of countries signs an FTA even if their industrialization levels are quite different (see the Concluding Remarks for more details).

2. The Model

2.1 Overview

Let $N = \{1, 2, \ldots, n\}$ be the set of n countries ($n \geq 3$), each of which is populated by a continuum of identical consumers who consume a numeraire good and a continuum of horizontally differentiated commodities that are indexed by $\omega \in [0,1]$. A differentiated commodity can be considered as a variety of an industrial good. Each industrial commodity ω is produced by one firm, also indexed by the same ω, which engages in price competition with other firms in individual segmented countries. We assume that there is no entry of firms into this industry. Each firm is owned equally by all domestic consumers who receive equal shares of all firms' profits. The numeraire good is produced competitively, on the other hand. Each consumer is endowed with l units of labor, which is used for production of the industrial and numeraire goods. Each unit of labor produces one unit of the numeraire good, so that the wage rate equals 1. We also assume that industrial commodities are produced with a constant-returns-to-scale technology, and normalize the unit labor requirement to be equal to 0 for each industrial commodity, without loss of generality. Alternatively, we can interpret the model such that each consumer is endowed with l units of the numeraire good, which can be transformed by a linear technology into industrial commodities.

In country $i \in N$, measure μ^i of consumers and measure s^i of firms that produce industrial commodities are located. Thus, country i produces s^i industrial commodities, which are consumed in every country in the world. Since the markets are segmented, firms can perfectly price discriminate among different countries. We normalize the size of total population so that $\sum_{k=1}^{n} \mu^k = 1$ as well as $\sum_{k=1}^{n} s^k = 1$. The ratio $\theta^i \equiv s^i/\mu^i$ measures country i's industrialization level. The higher the ratio, the higher the country's industrialization level. This ratio plays an important role later in our analysis. Country i imposes a specific tariff at a rate of t_j^i on the imports of the industrial commodities that are produced in country j. Under the Most-Favored-Nation (MFN) principle, country i must impose the same tariff rate against all other countries unless they are FTA partners of country i. We assume that there is no commodity tax, so that $t_i^i = 0$, and that the countries do not impose tariffs on the numeraire good, which may be traded internationally to balance trade. Tariff revenue is redistributed equally to domestic consumers.

2.1 Consumer demands

A representative consumer's utility is given by the following quasi-linear utility function:

$$U(q,q_0)=\int_0^1 q(\omega)d\omega-\frac{1-\sigma}{2}\int_0^1 q(\omega)^2 d\omega-\frac{\sigma}{2}\left[\int_0^1 q(\omega)d\omega\right]^2+q_0, \tag{1}$$

where q: $[0,1]\rightarrow\Re_+$ is an integrable consumption function, and q_0 denotes the consumption level of the numeraire good.[12] The second last term represents the substitutability among differentiated commodities, which may become clearer if we notice $\left[\int_0^1 q(\omega)d\omega\right]^2=\int_0^1\int_0^1 q(\omega)q(\omega')d\omega' d\omega$. The higher the parameter σ, the higher the substitutability between industrial commodities. The industrial commodities are independent from one another if $\sigma=0$, while they are perfect substitutes if $\sigma=1$. Letting y denote the consumer's income, the budget constraint can be written as

$$y=\int_0^1 \tilde{p}(\omega)q(\omega)d\omega+q_0, \tag{2}$$

where $\tilde{p}$: $[0,1]\rightarrow\Re_+$ denotes the consumer price function. The first order condition for the consumer's maximization problem gives us the inverse demand function for each good ω:

$$\tilde{p}(\omega)=1-(1-\sigma)q(\omega)-\sigma\int_0^1 q(\omega')d\omega'.$$

Integrating over $[0,1]$, we obtain

$$\int_0^1 q(\omega)d\omega=1-\tilde{P},$$

where $\tilde{P}=\int_0^1\tilde{p}(\omega)d\omega$ Substituting this equation back into the first order condition, we have

$$q(\omega)=\frac{1}{1-\sigma}\left[1-\tilde{p}(\omega)-\sigma(1-\tilde{P})\right].$$

2.3 Equilibrium in country i

Letting $p^i(\omega)$ and $\tilde{P}^i$ denote the producer price for commodity ω sold in country i,

[12] This utility function is a continuous-goods version of the ones of Shubik (1984) and Yi (1996, 2000) who analyze the case where there are only finitely many differentiated commodities. Our setup of continuous commodities is based on the model developed by Ottaviano et al. (2002). This specification is more suitable, for example, than perfectly competitive models for the analysis of FTA formation among asymmetric countries with a differentiated good, in which substitutability among differentiated commodities plays an important role. Interestingly, price competition and quantity competition yield the same equilibrium outcomes in this setup of continuous commodities since a firm's choice of either price or production quantity has only a negligible impact on the demands for other firms' products. Therefore, the following analysis would not be affected by the choice of strategic variables, which is another appealing feature of the model.

and the average consumer price in country i, respectively, a representative consumer's demands in country i for commodity ω produced in country k can be written as

$$q^i(\omega) = \frac{1}{1-\sigma}\left[1 - p^i(\omega) - t_k^i - \sigma(1-\tilde{P}^i)\right]. \tag{3}$$

The firm ω in country k chooses $\{p^i(\omega)\}_{i=1}^n$ to maximize its profits $\pi(\omega) = \sum_{i=1}^n \mu^i p^i(\omega) q^i(\omega)$. The first order condition for this maximization gives us

$$p^i(\omega) = \tfrac{1}{2}\left[1 - t_k^i - \sigma(1-\tilde{P}^i)\right], \tag{4}$$

for any i. Notice that $p^i(\omega)$ does not vary with ω. Prices charged by firms depend only on the importing country's tariff policies. We henceforth suppress the argument ω.

It follows from (4) that country i's average consumer price is

$$\begin{aligned}\tilde{P}^i &= \sum_{k=1}^n s^k (p^i + t_k^i)\\ &= \tfrac{1}{2}\left[1 + \bar{t}^i - \sigma(1-\tilde{P}^i)\right],\end{aligned}$$

Where $\bar{t}^i \equiv \sum_{k=1}^n s^k t_k^i$. Thus, country i's average consumer price $\tilde{P}^i$ is given by

$$\tilde{P}^i = \frac{1-\sigma+\bar{t}^i}{2-\sigma}. \tag{5}$$

Substituting (5) into (4) yields the equilibrium producer price p_k^i that each firm in country k charges for the market of country i, as a function of country i's tariff vector $\mathbf{t}^i = (t_1^i, \ldots, t_n^i)$:

$$p_k^i(\mathbf{t}^i) = \frac{1-\sigma}{2-\sigma} - \frac{1}{2} t_k^i + \frac{\sigma}{2(2-\sigma)} \bar{t}^i.$$

Then it follows from (3) that a representative consumer's demand in country i for a commodity produced in country k, denoted by q_k^i, is

$$q_k^i(\mathbf{t}^i) = \frac{1}{2-\sigma} - \frac{1}{2(1-\sigma)} t_k^i + \frac{\sigma}{2(1-\sigma)(2-\sigma)} \bar{t}^i. \tag{6}$$

Notice that $p_k^i(\mathbf{t}^i) = (1-\sigma) q_k^i(\mathbf{t}^i)$ for any tariff vector $\mathbf{t}^i$.

2.4 Social welfare

Under the world tariff vector t = (t^1, ..., t^n), each firm in country *i* earns the profits:

$$\pi_1(\mathbf{t}) = \sum_{k=1}^{n} \mu^k p_i^k(\mathbf{t}^k) q_i^k(\mathbf{t}^k) = \sum_{k=1}^{n} \mu^k (1-\sigma) q_i^k(\mathbf{t}^k)^2. \tag{7}$$

Country *i*'s *per capita* tariff revenue is

$$T^i(\mathbf{t}^i) = \sum_{k=1}^{n} t_k^i s^k\, q_k^i(\mathbf{t}^i). \tag{8}$$

A representative consumer's income in country *i* is the sum of labor income, redistributed tariff revenue, and the profit shares of the firms in country *i*:

$$y = l + T^i(\mathbf{t}^i) + \frac{s^i \pi_i(\mathbf{t})}{\mu^i} \tag{9}$$

Then it follows from (2) that

$$\begin{aligned}
q_0^i(\mathbf{t}) &= l + T^i(\mathbf{t}^i) + \frac{s^i \pi_i(\mathbf{t})}{\mu^i} - \sum_{k=1}^{n} s^k [p_k^i(\mathbf{t}^i) + t_k^i] q_k^i(\mathbf{t}_k^i) \\
&= l + \sum_{k=1}^{n} s^k t_k^i\, q_k^i(\mathbf{t}^i) + \frac{s^i}{\mu^i} \sum_{k=1}^{n} \mu^k p_i^k(\mathbf{t}^k) q_i^k(\mathbf{t}^k) - \sum_{k=1}^{n} s^k [p_k^i(\mathbf{t}^i) + t_k^i] q_k^i(\mathbf{t}_k^i) \\
&= l - \sum_{k \neq i} s^k p_k^i(\mathbf{t}^i) q_k^i(\mathbf{t}^i) + \frac{s^i}{\mu^i} \sum_{k \neq i} \mu^k p_{i,}^k(\mathbf{t}^k) q_i^k(\mathbf{t}^k),
\end{aligned} \tag{10}$$

where $q^i(\omega) = q_k^i(\mathbf{t}^i)$ if ω is produced in country *k*.

Substituting these equilibrium demands, (6) and (10), into (1), we obtain a representative consumer's utility in country *i* as a function of the world tariff vector, which can be considered as country *i*'s *per capita* social welfare:

$$W^i(\mathbf{t}) \equiv U((q_k^i(\mathbf{t}^i))_{k \in N}, q_0^i(\mathbf{t})) = V^i(\mathbf{t}^i) + X^i(\mathbf{t}^{-i}) - M^i(\mathbf{t}^i), \tag{11}$$

where

$$V^i(\mathbf{t}^i) \equiv U((q_k^i(\mathbf{t}^i))_{k \in N}, l), \tag{12}$$

$$M^i(\mathbf{t}^i) \equiv \sum_{k \neq i} s^k p_k^i(\mathbf{t}^i) q_k^i(\mathbf{t}^i) = \sum_{k \neq i} (1-\sigma) s^k\, q_k^i(\mathbf{t}^i)^2, \tag{13}$$

$$X^i(\mathbf{t}^{-i}) \equiv \frac{s^i}{\mu^i} \sum_{k \neq i} \mu^k p_i^k(\mathbf{t}^k) q_i^k(\mathbf{t}^k) = \frac{s^i}{\mu^i} \sum_{k \neq i} (1-\sigma) \mu^k q_i^k(\mathbf{t}^k)^2, \tag{14}$$

with $\mathbf{t}^{-i}=(\mathbf{t}^1, ..., \mathbf{t}^{i-1}, \mathbf{t}^{i+1}, ..., \mathbf{t}^n)$. The functions $V^i(\mathbf{t}^i)$, $M^i(\mathbf{t}^i)$, and $X^i(\mathbf{t}^{-i})$ represent a consumer's gross utility, import payments, and the export value of industrial commodities, respectively.[13] Country *i*'s social welfare consists of a consumer's gross utility $V^i(\mathbf{t}^i)$ and the per-capita industrial trade surplus $X^i(\mathbf{t}^{-i}) - M^i(\mathbf{t}^i)$.[14] Country *i*'s tariffs affect social welfare through the effects on $V^i(\mathbf{t}^i)$ and $M^i(\mathbf{t}^i)$. Other countries' tariffs affect country *i*'s social welfare through the effect on $X^i(\mathbf{t}^{-i})$.

Now, we examine the effects of tariff changes on the three components of social welfare: $V^i(\mathbf{t}^i)$, $X^i(\mathbf{t}^{-i})$, and $M^i(\mathbf{t}^i)$. We notice from (12)-(14) that an increase in a tariff rate affects these components only through the changes in the consumption of industrial commodities. We see from (6) that the consumption of an industrial commodity depends on the tariff rate imposed on that commodity and the average tariff rate, i.e., $q^i_k(\mathbf{t}^i) \equiv \tilde{q}^i_k(t^i_k, \bar{t}^i)$. Thus, we can write, for example, $V^i(t^i) = \tilde{V}^i(\tilde{q}^i_1(t^i_1, \bar{t}^i), ..., \tilde{q}^i_n(t^i_n, \bar{t}^i))$. An increase in t^i_j does not only affect q^i_j directly, but also affects q^i_k indirectly, for all $k = 1,2,\ldots,n$. These changes in consumption affect $V^i(\mathbf{t}^i)$ and $M^i(\mathbf{t}^i)$, in turn. As for the effect on $V^i(\mathbf{t}^i)$, for example, we have

$$\frac{\partial V^i}{\partial t^i_j} = \sum_{k=1}^{n} \frac{\partial \tilde{V}^i}{\partial \tilde{q}^i_k} \left(\frac{\partial \tilde{q}^i_k}{\partial t^i_j} + \frac{\partial \tilde{q}^i_k}{\partial \bar{t}^i} \frac{\partial \bar{t}^i}{\partial t^i_j} \right).$$

An increase in another country's tariff rate on country *i*'s commodity affects the export profits $X^i(\mathbf{t}^{-i})$ in a similar fashion. We can easily obtain the following lemma that shows the effects of raising a tariff rate on the three components of social welfare. The proof is straightforward and hence omitted.

Lemma 1. *The first order effects of raising t^i_j on V^i and M^i and the effect of raising t^j_i on X^i are:*

$$\frac{\partial V^i}{\partial t^i_j} = s^j \left[-\frac{1}{2-\sigma} + \frac{\sigma}{2(2-\sigma)} \sum_{k=1}^{n} s^k q^i_k(\mathbf{t}^i) + \frac{1}{2} q^i_j(\mathbf{t}^i) \right],$$

$$\frac{\partial X^i}{\partial t^i_j} = -\frac{\mu^j s^i q^j_i(\mathbf{t}^j)}{\mu^i} \left(1 - \frac{\sigma s^i}{2-\sigma} \right),$$

$$\frac{\partial M^i}{\partial t^i_j} = s^j \left[-q^i_j(\mathbf{t}^i) \left(1 - \frac{\sigma s^j}{2-\sigma} \right) + \sum_{k \neq i,j} q^i_k(\mathbf{t}^i) \frac{\sigma s^k}{2-\sigma} \right].$$

[13] The gross utility $V^i(\mathbf{t}^i) = U((q^i_k(\mathbf{t}^i))_{\{k \in N\}}, l)$ includes the utility derived from the consumption of l units of the numeraire good. However, since l is a constant that does not necessarily represents the actual consumption level of the numeraire good, $V^i(\mathbf{t}^i)$ should be regarded as the function that represents the gross utility derived from the consumption of the industrial commodities.

[14] This decomposition of social welfare, developed by Furusawa and Konishi (2004), may appear to suggest that a rise in industrial trade surplus unambiguously enhances social welfare. It should be emphasized, however, that the decomposition would not support such mercantilism, since an increase in imports, for example, is not necessarily bad as it raises consumers' gross utilities as well as it lowers trade surplus.

It may appear that an increase in a tariff rate of country i, say t^i_j , necessarily decreases the domestic consumer's gross utility V^i. Each consumer in country i reduces the consumption of country j's commodities as a consequence, which is detrimental. However, each agent consumes other commodities more than before, which tends to increase the consumer's gross utility. The latter indirect effect may outweigh the former so that an increase in a tariff rate may increase the domestic consumer's gross utility, if the industrial commodities are highly substitutable among themselves. Similarly, an increase in a tariff rate may not always decrease the import payments. If the industrial commodities are highly substitutable, the resulting decrease in q^i_j may be outweighed by increases in q^i_k for $k \neq i, j$. However, it is easy to see from Lemma 1 that an increase in another country's tariff unambiguously decreases the domestic profits obtained from the export to that country.

3. Free Trade Agreements

3.1 Incentives to sign an FTA

We examine incentives for country i to sign an FTA with country j. If countries i and j sign an FTA, they eliminate all tariffs imposed on commodities imported from each other, while keeping all other tariffs at their original levels. Letting $\mathbf{t}$ and $\mathbf{t}'$ denote the world tariff vectors before and after the FTA, respectively, $\mathbf{t}'$ is different from $\mathbf{t}$ only in the respect that $t^{i}_{j}{}' = t^{j}_{i}{}' = 0$. Country i has an incentive to sign an FTA with country j if and only if $W^i(\mathbf{t}') \geq W^i(\mathbf{t}')$, which can be written as

$$\Delta\, V^i(\mathbf{t}^i) + [\Delta X^i(\mathbf{t}^{-i}) - \Delta\, M^i(\mathbf{t}^i)] \geq 0, \qquad (15)$$

where Δ represents a change in the respective function values caused by an FTA between countries i and j such that $\Delta V^i(\mathbf{t}^i) \equiv \Delta V^i(\mathbf{t}^{i\prime}\,) - V^i(\mathbf{t}^i)$, for example. As we will see shortly, a tariff reduction is likely to increase a consumer's gross utility, unless the industrial commodities are highly substitutable from one another. Since the FTA increases country i's export profits and is also likely to increase the import payments, on the other hand, the FTA has an ambiguous impact on country i's industrial trade surplus. Under the MFN principle, each country i imposes the same external tariff rate, denoted by t^i, on all commodities imported from countries that have no FTAs with country i. We define $C_i = \{k \in N \,|\, t^i_k = 0\}$ as the set of countries that produce commodities on which country i does not impose tariffs. (Notice that C_i includes country i itself since $t^i_i = 0$.)

First, we investigate the sign of $\Delta V^i(\mathbf{t}^i)$. The next lemma shows that an FTA increases a consumer's gross utility of a country that has liberalized trade with the majority of countries, i.e., the majority of commodities are exempt from tariffs.

Lemma 2. *A bilateral FTA with country* j *increases a consumer's gross utility for country* i, *i.e.,* $\Delta\ V^i(\mathbf{t}^i) > 0$, *if* $s^{C_i} + (s^j/2) \geq ½$.

Remark 1. The condition reflects the second best effect: In an economy with distortions, the partial removal of tax distortions may reduce efficiency. When a tariff on a commodity is eliminated, distortions between this commodity and untaxed commodities shrink, whereas distortions with taxed commodities expand. Thus, if there are more untaxed commodities than taxed commodities, the second best theory tells us that a bilateral FTA between i and j is likely to raise a consumer's gross utility. The condition $s^{C_i} + (s^j/2) \geq ½$ matches exactly to this observation.

Next, we turn to investigating the effect of an FTA between countries *i* and *j* on the industrial trade surplus. Let M^i_k and X^i_k be country *i*'s (per capita) import payments to country *k* and country *i*'s (per capita) export profits from country *k*, respectively:

$$M^i_k(\mathbf{t}^i) = (1-\sigma)s^k q^i_k(\mathbf{t}^i)^2 \tag{16}$$

$$X^i_k(\mathbf{t}^k) = \frac{s^i}{\mu^i}(1-\sigma)\mu^k q^k_i(\mathbf{t}^k)^2 \left(= \frac{\mu^k}{\mu^i} M^k_i(\mathbf{t}^k) \right) \tag{17}$$

Then, we can rewrite country *i*'s industrial trade surplus as

$$X^i(\mathbf{t}^{-i}) - M^i(\mathbf{t}^i) = \sum_{k \neq i} [X^i_k(\mathbf{t}^k) - M^i_k(\mathbf{t}^i)]$$

An FTA between *i* and *j* only involves changes in $\mathbf{t}^i$ and $\mathbf{t}^j$ so that it does not affect $X^i_k(\mathbf{t}^k)$ for any $k \neq i, j$. Consequently, a change in country *i*'s industrial trade surplus can be written as

$$\Delta[X^i(\mathbf{t}^{-i}) - M^i(\mathbf{t}^i)] = \underbrace{\Delta[X^i_j(\mathbf{t}^j) - M^i_j(\mathbf{t}^i)]}_{\text{Direct surplus effect}} \underbrace{- \sum_{k \neq i,j} \Delta M^i_k(\mathbf{t}^i)}_{\text{Third country effect}}$$

The third country effect, represented by the last terms, is always positive since the reduction of t^j_i makes commodities imported from country *j* relatively less expensive, and hence country *i*'s imports from third countries decrease, i.e., $\Delta M^i_j(\mathbf{t}^i) < 0$. The reduction of FTA signatories' imports from all other countries hurts those outsiders, but provides countries *i* and *j* with incentives to sign an FTA.

Having shown that the third country effect is positive, let us now investigate the direct surplus effect, which can be rewritten as follows from (16) and (17):

$$\Delta[X_j^i(\mathbf{t}^j) + M_j^i(\mathbf{t}^i)] = \mu^j(1-\sigma)\,\Delta[\theta^i q_i^j(\mathbf{t}^j)^2 - \theta^j q_j^i(\mathbf{t}^i)^2,]$$

where $\theta^i = s^i/\mu^i$ as defined above. The higher θ^i and the lower θ^j, the larger an increase in country *i*'s industrial trade surplus. Thus, the direct surplus effect is unbalanced in favor of the relatively more industrialized country.[15] The more industrialized country derives a large benefit from the opening of the partner's relatively large market. In addition, opening its own market to the partner's firms does not significantly increase import payments since the resulting penetration by the partner's firms is relatively small. Another important factor that affects the incentives to form an FTA is the difference in the original tariff rates. If $t^i < t^j$, for example, then it is likely that $\Delta q_j^i(\mathbf{t}^i) < \Delta q_i^j(\mathbf{t}^j)$. Country *i*'s export to country *j* increases more than its import from country *j*, and hence the FTA between *i* and *j* tends to be more beneficial to country *i*.

3.2 Stable free trade networks

An FTA that involves more than two countries can be considered as a collection of bilateral FTAs between member countries, so in the graph theory an arbitrary network of FTAs can be described as a graph. An FTA between countries *i* and *j* can be considered as a *link*, which is an unordered pair of two countries. An *FTA graph* is an undirected graph, (N, Γ), consisting of the set of countries N and a (free trade) *network* Γ that is a collection of links. The set of country *i*'s FTA *partners* in network Γ is $C_i(\Gamma) = \{i\} \cup \{k \in N : (i, k) \in \Gamma\}$, which includes *i*, as we have already described. We continue to write C_i without confusion, as long as network Γ is fixed.

If external tariff rates are exogenously determined as in this paper, or if they are determined uniquely for each free trade network Γ (such as in the case where all countries set their individual optimal tariffs given a prevailing network Γ), then country *i*'s payoff (social welfare) can be written uniquely by $u_i(\Gamma)$. The set of countries N and their payoff functions define a *network formation game*.

Network formation games are first studied by Jackson and Wolinsky (1996). A *pairwise stable network* is a network Γ^* such that (i) for any $i \in N$ and for any $(i, j) \in \Gamma^*$, $u_i(\Gamma^*) \geq u_i(\Gamma^* \backslash (i, j))$, i.e., no country has an incentive to cut a link with another, and (ii) for any $(i, j) \notin \Gamma^*$ with $i \neq j$, if $u_i(\Gamma^*) < u_i(\Gamma^* \cup (i, j))$ then $u_j(\Gamma^*) > u_j(\Gamma^* \cup (i, j))$, i.e., for any unlinked pair of countries, at least one of them has no incentive to form a link with the other.[16]

[15] Indeed, if one country's direct surplus effect is positive, the partner's direct surplus effect must be negative since the sum of two countries' direct surplus effects is always zero, i.e., $\Delta X_j^i(\mathbf{t}^j) = \Delta M_i^j(\mathbf{t}^j)$ for any $i, j \in N$ with $i \neq j$.

[16] Readers may be tempted to formulate a strategic form game such that each player (country) announces the names of players with whom she wants to be linked, and a link is formed if and only if both sides of the link announce each other's names. In such a game, however, there would be too many Nash equilibria, always including the one without any link. It is because a player has no incentive to announce the name of

We are particularly interested in the situation where global free trade is effectively attained. A *complete* graph is the graph $(N, \Gamma^{\text{comp}})$ that contains all possible links, i.e., for any $i, j \in N$ with $i \neq j$, $(i, j) \in \Gamma^{\text{comp}}$. We call Γ^{comp} a *complete network*. The global free trade is a complete graph in the free trade network formation game.

3.3 Symmetric countries

We say that countries i and j are *symmetric* if $s^i = s^j$ and $\mu^i = \mu^j$. This subsection considers the case in which the world consists of n symmetric countries so that $s^i = \mu^i = 1/n$ for any $i \in N$. In this case, country i's direct surplus effect can be simplified as

$$\Delta[X_j^i(\mathbf{t}^{-i}) - M_j^i(\mathbf{t}^i)] = \mu^j(1-\sigma)\Delta\left[\frac{s^i}{\mu^i} q_i^j(\mathbf{t}^j)^2 - \frac{s^i}{\mu^j} q_j^i(\mathbf{t}^i)^2\right]$$
$$= \frac{1-\sigma}{n}[\Delta(q_i^j(\mathbf{t}^j)^2) - \Delta(q_j^i(\mathbf{t}^i)^2)].$$

The current network structure affects the impact of the FTA between i and j on country i's industrial trade surplus through its effects on commodity demands. Especially important is the size of C_i and C_j.

Let us say that countries i and j are *completely symmetric* if they are symmetric and $|C_i| = |C_j|$. If the original tariffs are the same between completely symmetric countries i and j, i.e., $t^i = t^j = t$, then $\bar{t}^i = \bar{t}^j$ and $q_i^j(\mathbf{t}^j) = q_j^i(\mathbf{t}^i)$, and hence we have $\Delta q_i^j(\mathbf{t}^j) = \Delta q_j^i(\mathbf{t}^i)$ and $\Delta[X_j^i(\mathbf{t}^{-i}) - M_j^i(\mathbf{t}^i)]$. Thus, the direct surplus effect disappears if countries i and j are completely symmetric and their original tariffs are the same. An increase in country i's export to country j and an increase in country i's import from county j are completely canceled out. On the other hand, the third country effect is nonnegative. Thus, we have $\Delta[X^i(\mathbf{t}^{-i}) - M^i(\mathbf{t}^i)] \geq 0$ if countries i and j are completely symmetric.

Completely symmetric countries always have incentives to sign an FTA as long as the condition in Lemma 2 is satisfied. One important case is that all pairs but (i, j) have already formed free trade links. Since most tariffs are already eliminated, an FTA between i and j reduces distortions, and hence enhances a consumer's gross utility in these countries ($\Delta V^i > 0$). Thus, the two countries can improve social welfare by signing an FTA, which leads to our first proposition.[17]

the player who does not announce her name. See Dutta and Mutuswami (1997) for the coalition-proof Nash equilibrium, a refinement of the Nash equilibrium in such games

[17] Bagwell and Staiger (1999a) argue that reciprocal trade liberalization between two countries is beneficial to both countries since it leaves each country's terms of trade unchanged so that it eliminates negative terms-of-trade externalities. An FTA between two completely symmetric countries fits their argument in that it leaves the bilateral (industrial commodity) terms of trade unaffected. In addition, each country's bilateral terms of trade against a third country improves as $q_k^i(\mathbf{t}^i)$ and hence $p_k^i(\mathbf{t}^i)$ declines for $k \neq i, j$.

Proposition 1. *Suppose that there are n symmetric countries in the world, and that their external tariff rates are the same if they are imposed. Then, global free trade (the complete network* Γ^{comp} *is a stable network.*

Proof. The second condition for pairwise stability is vacuously satisfied since there is no unlinked pair of countries under the complete free trade network. Therefore, we need only show that a representative country i has no incentive to cut a link with country j. Or equivalently, country i has an incentive to sign an FTA with country j under the network $\Gamma^{\mathrm{comp}}\backslash(i, j)$. Now, we know from the above observation that country i's industrial trade surplus does not decrease by signing the FTA since countries i and j are completely symmetric. Moreover, since $s^{Ci} = 1 - (1/n)$ and $s^{j} = 1/n$, we have $s^{Ci} + (s^{j}/2) = 1 - (½n) > ½$ for all $n \geq 3$ under $\Gamma^{\mathrm{comp}}\backslash(i, j)$. Then, it follows from Lemma 2 that a consumer's gross utility in country i strictly increases. Therefore, we have $u_i(\Gamma^{\mathrm{comp}}) > u_i(\Gamma^{\mathrm{comp}}\backslash(i, j))$, implying that Γ^{comp} is a stable network. ■

Remark 2. Note that this proposition holds even in the case where each country optimally adjusts its tariff rate to a change in the free trade network. If a country cuts a link with another under Γ^{comp}, these countries would impose the same optimal tariff rate by symmetry. Thus, the assumptions of Proposition 1 are satisfied even if tariff rates are endogenously determined at their optimal levels.

Bagwell and Staiger (2005) show that any Pareto efficient tariff vector is unstable since a pair of countries can benefit from reciprocal reduction of their tariffs against each other while retaining those against other countries. This bilateral opportunism problem arises since the mutual tariff reduction that is discriminatory against third countries will improve their terms of trade against third countries. Their striking proposition also holds in our imperfectly competitive world. The bilateral tariff reduction from a Pareto efficient tariff vector can be tailored so as to nullify the direct surplus effect. Since the third country effect is always positive, however, this tariff reduction will unambiguously improve the industrial trade surplus, so any Pareto efficient tariff vector is vulnerable to the bilateral opportunism. Due to the imperfectness of competition, free trade tariff vector in our model (the origin of the tariff space) lies above the set of Pareto efficient tariff vectors. Therefore, bilateral tariff reduction from free trade, i.e., mutual provision of import subsidies, definitely benefits both countries, implying that free trade is not pairwise stable if a pair of countries can choose discriminatory subsidies when they sign an FTA. Although we restrict the feasible set of tariff vectors to the non-negative orthant following the convention of the literature, allowing countries to choose subsidies can be an interesting extension of our analysis of FTA network formation game.

Now, it is natural to ask if the complete graph is a unique stable network. Unfortunately, it is not the case in general even if countries are symmetric. If $q_i^j(\mathbf{t}^j)$ is significantly smaller than $q_j^i(\mathbf{t}^i)$ and hence $\Delta q_i^j(\mathbf{t}^j)$ is significantly smaller than $\Delta q_j^i(\mathbf{t}^i)$, the direct surplus effect for country *i* is negative and it may outweigh the third country effect. This situation arises when country *j* has many FTAs with other countries, while country *i* has a small number of FTAs.

Lemma 3. *Consider the case where the world consists of n symmetric countries that would set a common tariff rate of t. Country i's incentive to sign an FTA with country j increases with* $|C_i|$ *and decreases with* $|C_j|$*, and hence it is smallest if country i does not belong to any FTA while country j has FTAs with all countries but i.*

Consider the situation where country *i*'s incentive to have an FTA with country *j* is smallest as described in Lemma 3. If σ is large and close to unity, consumer demands for a commodity are sensitive to prices for other commodities. In the absence of an FTA, therefore, isolated country *i* does not import much of industrial commodities, and most of industrial commodities consumed are domestically produced. However, once country i signs an FTA with country *j*, much of (about a half of) the consumption of domestic commodities is substituted by those produced in country *j* so that country *i* experiences a dramatic increase in its import payments. In contrast, country *j* has already opened its market to all but country *i* before the FTA. Therefore, the FTA with country *i* does not increase its imports much even if σ is large. Therefore, the direct surplus effect of country *i* is negative and large in magnitude, which outweighs the third country effect and the effect on $\Delta V^i(\mathbf{t}^i)$. Although it is hoped that (preferential) trade liberalization continues under the GATT Article XXIV, it is quite possible that the process of FTA formation stops prematurely even if all countries are symmetric.

Now, we seek the condition under which every pair of countries has incentive to form an FTA regardless of the current FTA network. In such a case, the complete network (global free trade) becomes a unique stable network. The next proposition states that the complete FTA network is a unique stable network if tariffs are small or if the industrial commodities are not highly substitutable from one another.

Proposition 2. *Suppose that the world consists of n symmetric countries that would set a common external tariff rate of t. Any pair of countries without an FTA have incentives to form a free trade link under any network Γ, and thus the complete FTA network* Γ^{comp} *is a unique stable network if and only if either*

(i) $A(\sigma,n) \equiv -4\mathrm{n} + 4(5n-8)\sigma - (11n - 23 + \frac{4}{n})\sigma^2 \le 0$, *or*

(ii) $t \le \tau(\sigma, n) \equiv \frac{8(1-\sigma)(n-2\sigma)}{A(\sigma,n)}$ *when* $A(\sigma, n) > 0$

is satisfied. Condition (i) is satisfied if σ is smaller than or equal to the smaller root of $A(\sigma,n)=0$, which we call $\sigma(n) \in (0,1)$. The critical tariff rate $\tau(\sigma, n)$ in condition (ii) is decreasing in σ.

Figure 1 depicts the threshold for the uniqueness of the stable network. The condition in Proposition 2 is satisfied if (σ, t) lies to the left of the graph of $\tau(\cdot, n)$. If this condition is violated, there exists a pairwise stable network, in addition to the complete FTA network, such that one country is isolated while all other countries have FTAs with one another. If $n = 15$ and $\sigma = .98$, for example, the condition in Proposition 2 is violated for $t = .04$. In such a situation, all but country 1, say, have FTAs with one another. This network is stable since the isolated country 1 does not have an incentive to have a bilateral FTA with any other country.

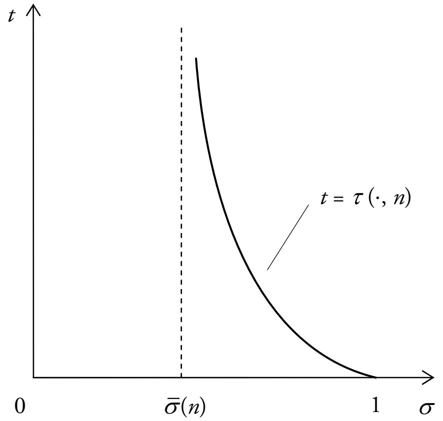


Figure 1. The region where the global free trade is a unique stable network

This proposition suggests that FTA formation and multilateral trade negotiation under the auspices of the WTO are complementary. As tariff rates decline through multilateral negotiations, it becomes more likely that unlinked pairs of countries have FTAs, leading to the complete network of FTAs.

Moreover, under the condition where Proposition 2 applies, the world free trade network will eventually reach the complete network such that global free trade is effectively attained if countries myopically make decisions as to whether or not they sign FTAs with other countries. For dynamic network formation games, Watts (2001) defines a stable state as the network in which any randomly selected pair of myopic players have no incentive to severe the link if they are currently linked and to form a link if they are not linked. The complete FTA network is the unique stable state if the condition of Proposition 2 is satisfied. This result can also be extended to the case of farsighted countries with an arbitrary discount rate. Applying Theorem 3 of Dutta et al. (2005), we can conclude that if the condition of Proposition 2 is satisfied, there is a Markov perfect equilibrium in which the complete FTA network is eventually reached from any FTA network.

3.4 Asymmetric countries

Let us turn to a more realistic case in which countries are asymmetric. As we infer from the preceding analysis, countries are less likely to have FTAs and the complete FTA network is less likely to be pairwise stable in an asymmetric world. Of course, a pair of countries with similar size of the market and industry still signs a bilateral FTA. Moreover, we show in this subsection that countries with similar industrialization levels, but not necessarily similar in the absolute size of the market and industry, tend to sign a bilateral FTA. To this end, we assume here that $\sigma = 0$. Although this simplification is restrictive, it highlights how the asymmetry of countries affects the FTA network formation.

In this special case of no substitution among industrial commodities, we can easily calculate social welfare of each country. Since commodity demands are independent of one another when $\sigma = 0$, the main part of a consumer's gross utility can be written as a simple sum of utilities derived from the consumption of all individual commodities. Let $p(t)$ and $q(t)$ denote the equilibrium producer price and quantity of the industrial commodity that is faced with the tariff rate t, and let $v(t)$ denote a consumer's utility derived from the consumption of that commodity. Then, we can write

$$V^i(\mathbf{t}^i) = \sum_{k\in C_i} s^k v(0) + \sum_{h\notin C_i} s^h v(t) + l,$$

$$X^i(\mathbf{t}^{-i}) = \frac{s^i}{\mu^i}\left[\sum_{k\in C_i\setminus\{i\}} \mu^k p(0)q(0) + \sum_{h\notin C_i} \mu^h p(t^h)q(t^h)\right],$$

$$M^i(\mathbf{t}^i) = \sum_{k\in C_i\setminus\{i\}} s^k p(0)q(0) + \sum_{h\notin C_i} s^h p(t^h)q(t^h).$$

As Figure 2 shows we have, for $t \leq 1$, that

$$p(t) = \frac{1+t}{2} - t = \frac{1-t}{2},$$

$$q(t) = \frac{1-t}{2},$$

and hence

$$v(t) = \frac{(1-t)(3+t)}{8},$$

$$p(t)q(t) = \frac{(1-t)^2}{4}.$$

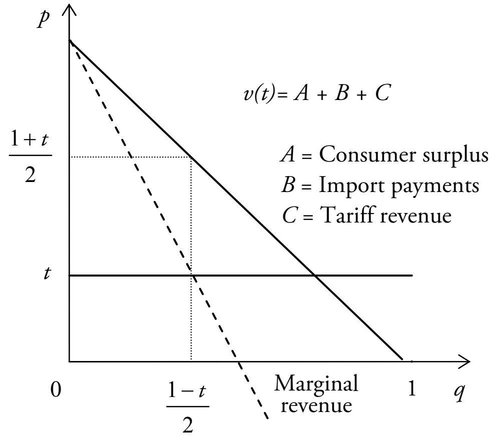


Figure 2. Equilibrium in a commodity market

If countries i and j sign an FTA, then C_i expands to include j. Thus, the impact on country i's welfare is

$$\begin{aligned}\Delta W^i(\mathbf{t}) &= s^j[v(0)-v(t^i)]+\frac{s^i\mu^j}{\mu^i}[p(0)q(0)-p(t^j)q(t^j)]-s^j[p(0)q(0)-p(t^i)q(t^i)]\\ &=\mu^j\left[\frac{\theta^j t^i(2+t^i)}{8}+\frac{\theta^i t^j(2-t^j)}{4}-\frac{\theta^j t^i(2-t^i)}{4}\right]\\ &=\frac{\mu^j}{8}[\theta^j t^i(3t^i-2)+2\theta^i t^j(2-t^j)] \qquad (18)\end{aligned}$$

The first observations we derive from (18) are rather obvious. Excluding prohibitive tariffs from consideration, we find that the higher is t^j the higher is $\Delta W^i(\mathbf{t})$. Country i benefits more from the FTA with country j as country j's original tariff rate is high. As for country i's own tariff, we should distinguish between two cases, whether or not t^i is smaller than the optimal tariff 1/3. If $t^i \le 1/3$, the lower is t^i, the higher is $\Delta W^i(\mathbf{t})$. If $t^i > 1/3$, on the other hand, the opposite is true. If t^i is higher than the optimal tariff for some reason, country i has an incentive to unilaterally cut its tariff at least to the optimal level. This incentive becomes greater as t^i increases. Indeed, as (18) indicates, $\Delta W^i(\mathbf{t})$ is unambiguously positive if $t^i > 2/3$. Henceforth, we restrict our attention to the case where $t^i \le 1/3$ for any $i \in N$, as no country has an incentive to select a higher tariff rate than its optimal level.

How do the countries' industrialization levels affect country i's incentive to sign the FTA? It follows from (18) that country i has an incentive to sign the FTA with country j if and only if

$$\frac{\theta^j}{\theta^i}\le\frac{2t^j(2-t^j)}{t^i(2-3t^i)}. \qquad (19)$$

Country i benefits from the FTA with country j if country j's industrialization level, relative to its own, is not so large. FTAs are reciprocal concessions: Each signatory gives up exercising its market power in import good markets in exchange for obtaining better access to export good markets in its partner countries. Thus, it is intuitive that the FTA is more beneficial if the resulting increase in its export to the partner is large (i.e., s^i and μ^j are large) and the resulting increase in its import from the partner is small (i.e., s^j and μ^i are small). Changes in country i's export and import depend, in general, on the FTA configuration of countries j and i, respectively. In the current case of $\sigma = 0$, however, they hinge on the bilateral trade relationship between i and j. Gains from the FTA are the simple sum of individual gains across the varieties. If s s^i, μ^i, s^j, and μ^j are all doubled (so that θ^j/θ^i is unchanged), for example, the gains from the FTA are also doubled as (19) indicates, leaving the sign of $\Delta W^i(\mathbf{t})$ as it was.

In order for the FTA between countries i and j to be signed by both countries, the counterpart of (19) for country j must also be satisfied. Assuming $t^i = t^j \equiv t$ for clarity, we find that the FTA is signed if and only if

$$\frac{2-3t}{4-2t} \leq \frac{\theta^j}{\theta^i} \leq \frac{4-2t}{2-3t}. \tag{20}$$

As t increases from 0 to 1/3, this range of θ^j/θ^i expands from $[1/2, 2]$ to $[3/10, 10/3]$. The higher is t, the greater is the benefit of the FTA; hence even asymmetric countries sign an FTA. We record the finding for the case of $t^i = t^j$ in the following proposition.

Proposition 3. *Suppose that $\sigma = 0$ and that countries would impose the common tariff rate t as their external tariffs. Countries i and j form a link if their industrialization levels are similar such that (20) is satisfied. The stable network is a generically unique collection of all links, each of which connects such a pair of countries.*

If countries' industrialization levels are not too different, then they have incentives to form an FTA. Countries with similar industrialization levels tend to form a link since (i) each country wants to sign an FTA with a country whose industrialization level is not so large compared with its own and (ii) an FTA is put into force only if it is signed by both parties. Suppose that there are two groups of countries: one is a group of developed countries with similar and high industrialization levels, and the other is a group of less developed countries with similar and low industrialization levels. Suppose also that every country selects its external tariff at its optimal level 1/3 for concreteness. Then, if the industrialization level of each developed country is far greater (more than 10/3 times) than the one of any less developed country, the FTA formation process leads to a stable network in which all countries within each group are linked with each other, while there is no link across the two groups. The FTA formation process may end with two (stumbling) trading blocs if industrialization levels of two groups are very different from each other.

4. Free Trade Agreements vs. Customs Unions

This section investigates the difference in member countries' incentives to sign a new FTA emphasizing the fact that a CU requires that all members be involved when a member country wants to have a free trade link with an outside country. The main goal of the paper is to assess how far the process of PTAs continues and whether or not global free trade is effectively attained as a complete world-wide web of PTAs. The analysis in this section possibly tells us which form of PTAs, CU or FTA, should be encouraged for facilitating more PTAs in the world. In order to focus on the issue, we assume that external tariff rates are the same in both cases.

We compare country *i*'s incentives to have a new free trade link with country $j \notin C_i$ between two cases: the case where C_i forms a CU and the case where C_i is a *regional FTA* such that every pair of countries in C_i has a bilateral FTA. We begin with investigating the impact on a consumer's gross utility V^i. As Section 3.1 shows, the impact on V^i is ambiguous in both cases. However, these effects are exactly the same between the two cases, since V^i only depends on $\mathbf{t}^i$ and changes in $\mathbf{t}^i$ are the same between the two cases. Thus, the difference in changes of the industrial trade surplus between these two cases will determine whether or not country *i*'s incentive to have an FTA with country *j* is higher in the case where C_i is a CU rather than a regional FTA. Here, we decompose the third country effect into the member country and nonmember country effects:

$$\Delta X^i - \Delta M^i = \underbrace{(\Delta X^i_j - \Delta M^i_j)}_{\text{Direct surplus effect}} + \underbrace{\sum_{k \in C_i \setminus \{i\}} (\Delta X^i_k - \Delta M^i_k)}_{\text{Member country effect}} + \underbrace{\sum_{k \notin C_i \cup \{j\}} (\Delta X^i_h - \Delta M^i_h)}_{\text{Nonmember country effect}},$$

where country *k* is a representative partner of *i*, i.e., $k \in C_i \setminus \{i\}$, and country *h* is a representative outsider of *i*, i.e., $h \notin C_i \cup \{j\}$.

Table 1 depicts the signs of the effects, and compares these two cases item by item. Similarly to the impacts on V^i, the effects of an FTA with country *j* on $M^i_j = M^i_j(\mathbf{t}^i)$ are the same between the two cases, since country *i*'s imports from country *j* are solely determined by $\mathbf{t}^i$. This effect is positive since country *i* lowers its tariff rate for commodities imported from country *j*. In contrast, the effects on $X^i_j = X^i_j(\mathbf{t}^j)$ are different especially when $|C_i|$ is large. It is because country *j* eliminates tariffs against all countries in C_i in the case of CU while it eliminates tariffs only for commodities imported from country *i* in the case of FTA. Since industrial commodities are substitutable from one another, it is obvious that an increase X^i_j is smaller in the case of CU. Consequently, the direct surplus effect is smaller in the case of CU than in the case of FTA.

Table 1. FTA vs. CU ($\sigma > 0$)

	FTA		CU
ΔV^i	?	=	?
ΔM^i_j	+	=	+
ΔX^i_j	+	>	+
ΔM^i_k	–	=	–
ΔX^i_k	0	>	–
ΔM^i_h	–	=	–
ΔX^i_h	0	=	0

Next, we investigate the effects on country *i*'s industrial trade surplus with a member country $k \in C_i \backslash \{i\}$. As before, the effects on $M^i_k = M^i_k(\mathbf{t}^i)$ are the same in both cases. However, the effects on $X^i_k = X^i_k(\mathbf{t}^k)$ are different again. In the case of FTA, $\mathbf{t}^k$ is unaffected and hence X^i_k does not change. In the case of CU, on the other hand, country *k* also eliminates tariffs against country *j*, and country *i*'s export to country *k* is reduced due to the substitution effect. Country *i*'s industrial trade surplus with a member country *k* is again lower in the case of CU. Finally, it is easy to see that the third country effects with nonmembers are the same in both cases. Import payments to country *h* decrease by the same amount due to the tariff reduction for commodities imported from country *j*, and country *i*'s exports to country *h* stay the same in both cases since $\mathbf{t}^h$ is not affected.

We have shown that the impacts of a new FTA on a consumer's gross utility are the same between the two cases, but the changes in the industrial trade surplus is unambiguously smaller in the case of CU. We record this result as a lemma.

Lemma 4. *Country i has less incentive to have a free trade link with country $j \notin C_i$ when C_i forms a CU rather than a regional FTA, unless the industrial commodities are independent of one another, i.e., $\sigma = 0$, in which case the incentives are the same.*

Whether or not country *i*'s incentive to have a free trade link with country *j* is lower when C_j forms a CU rather than a regional FTA is generally ambiguous, however. The difference between these two cases in our terminology is that country *i* adds only one link with country *j* in the case of a regional FTA, whereas in the case of a CU country *i* adds $|C_j|$ links simultaneously with all individual countries in C_j. The latter case is effectively equivalent to the case where country *i* has an FTA with an integrated economy that consists of all countries in C_j. Whether country *i* prefers having a free trade link with country *j* alone or with the whole C_j depends on the relative characteristics of *j* and C_j.

However, we can make a strong statement in the case of symmetric countries with a low substitution parameter σ. Proposition 2 indicates that if all countries are symmetric and if σ is not very high, country i has an incentive to have an FTA with any country in any FTA configuration, in particular with country j alone or with all countries comprising C_j. Therefore, country i wants to have a free trade link with country j regardless of whether C_j forms a CU or an FTA. Combining this observation together with Lemma 4, we find that two countries are less likely to form a link if either of them is a member of a CU. Indeed, the complete FTA network is the unique stable network if all PTAs take a form of FTA (Proposition 2), whereas several CUs of asymmetric size may co-exist in a stable network if all FTAs take a form of CU.[18]

Proposition 4. *Suppose that countries are symmetric, imposing the same external tariff rate and that the condition in Proposition 2 is satisfied. Then, a pair of countries is less likely to have a free trade link if either of them is a member of a CU rather than a regional FTA.*

If countries are not symmetric, CUs can facilitate global trade liberalization more than FTAs. Consider again the case of asymmetric countries with $\sigma = 0$ in which every country would select its external tariff rate at its optimal level 1/3. We order n asymmetric countries according to their industrialization levels such that $\theta^1 \geq \theta^2 \geq \ldots \geq \theta^n$. Proposition 3 implies that if $\theta^1 / \theta^n > 10/3$, countries 1 and n will not sign an FTA, and the process of bilateral FTA formation will never reach global free trade. However, if all PTAs take a form of CU, the process of CU formation may reach global free trade. Let us consider a CU by $C(k) \equiv \{1,2, \ldots,k\}$, the set of k countries with highest industrialization levels. The industrialization level of the entire $C(k)$, i.e., $\theta^{C(k)} \equiv \sum_{h\in C(k)} s^h / \sum_{h\in C(k)} \mu^h$, is the 'average' industrialization level of all individual members of $C(k)$, so that $\theta^1 \geq \theta^{C(k)} \geq \theta^k$. Now, it follows from Proposition 3 that $C(k)$ and $k + 1$ sign an FTA, or in other words, CU by $C(k)$ expands to include $k + 1$, if $\theta^{C(k)} / \theta^{k+1} \leq 10/3$. Notice that this inequality can hold even if $\theta^1/\theta^{k+1} > 10/3$. The CU formation averages out member countries' industrialization levels, and hence encourages a less industrialized country to join the group. In particular, if $\theta^{C(k)}/\theta^{k+1} \leq 10/3$ for any $k = 1, \ldots, n-1$, CUs serve as 'building blocks' and the process of CU formation will reach global free trade.[19]

[18] Employing a coalition bargaining game, Yi (1996) shows that in equilibrium, two CUs of different size are formed when the world consists of a reasonable number of symmetric countries. We can conduct the same exercise in our model and obtain qualitatively the same result.

[19] We should note that history of CU expansion may matter. It is possible for the CU expansion to stop prematurely if two unions, one by developed countries and the other by less developed countries, are formed, and the difference in the industrialization levels of these two unions is quite large.

5. Concluding Remarks

We have introduced a general analytical framework that is suitable for the investigation of PTAs and shown how countries' incentives vary with the country size, industrialization level, substitutability among industrial commodities, etc. We have found that if all countries are symmetric, the complete FTA network is pairwise stable and it is the unique stable network if industrial commodities are not highly substitutable from one another or if predetermined external tariff rates that countries would choose are small. We have also compared FTAs and CUs as to which of these two regimes facilitates PTA formation. We have shown that in the symmetric country case where industrial commodities are not highly substitutable, countries are likely to have less incentive to have a new free trade link if one of the countries is a member of a CU rather than an FTA. If countries are asymmetric, however, CU formation averages out member countries' industrialization levels, which may help further CU formation.

The present paper introduces a model that fits the analysis of FTAs and derives some useful results that are summarized above. However, it is naturally far from a complete analysis of FTAs. We examine elsewhere (Furusawa and Konishi, 2005) FTA network formation when transfers between signatories are allowed. With transfers, a pair of countries signs an FTA if and only if the FTA enhances the joint social welfare. Since the third country effects are always positive and the sum of the direct surplus effects is zero regardless of the heterogeneity between the countries, they are quite likely to sign the FTA. Indeed, Propositions 1 and 2 in this paper can be generalized to the case of asymmetric countries. Although we obtain stronger results when transfers between FTA signatories are allowed, feasible amounts of transfer are usually limited in practice. Thus, both of this paper and the companion paper provide useful insights of the problem. As for a further extension in this direction, it may be interesting to consider more generalized forms of transfers such as subsidizing other links in a more general environment (see Bloch and Jackson, 2004).

Another obvious extension is to relax the assumption on the selection of external tariffs. We have assumed throughout the paper that external tariff rates are exogenously fixed, since it is necessary to simplify the model for analyzing various forms of complicated FTA networks. If we assume instead that countries always set optimal tariffs given their FTA link structures, then they have incentives to lower their external tariffs as they form more free trade links, which Bagwell and Staiger (1999b) call the tariff complementarity effect. Indeed, Richardson (1993), Bagwell and Staiger (1999b), Yi (2000), and Ornelas (2005a) demonstrate in their respective models that if FTA signatories optimally adjust their individual external tariffs, an FTA induces the signatories to cut their tariffs so deeply that their imports from nonmember countries increase, i.e., the nonmember country effect, which is part of the third country effect, is negative. It can be shown that the same result obtains in our model if we allow FTA signatories to optimally adjust their external tariff. Yi (2000) and Ornelas (2005c) further show that global free trade may not be realized due to the free rider problem caused by this tariff complementarity effect. A similar result is expected to obtain in our extended model, i.e., there may be an asymmetric incomplete stable FTA network such as only one country is isolated from the rest of the countries. Nevertheless, as Remark 2 indicates, the complete FTA network continues to be stable even if the external tariffs are optimally adjusted.

Moreover, Proposition 2 suggests that the complete FTA network may survive as a unique stable network as countries symmetrically expand their FTA network, lowering their external tariffs *symmetrically* in the process. Let us imagine a dynamic FTA formation such that in each step, all countries have the same number of FTA links. As the FTA formation proceeds, their external tariffs decline and eventually enter the region where the complete FTA network is a unique stable network when external tariffs are fixed (see Proposition 2). Consider a pair of countries that form a new FTA link in this phase of the FTA formation. Due to the symmetry, the direct surplus effect is nil. The third country effect may be negative as the nonmember country effect is negative as we have seen above. But in the phase where they have already formed several FTAs, the member country effect, which is positive as a decrease in the external tariff further reduces the import from member countries, is likely to outweigh the nonmember country effect so that the entire third country effect is positive. Indeed, our extensive numerical analysis, which is available upon request, indicates that every pair of completely symmetric countries has incentive to sign an FTA so that if all countries symmetrically expand their FTA networks, the FTA formation continues until the complete FTA network is reached even though the external tariffs are optimally adjusted in each step.

Introducing governments' political motivation to the model, such as Ornelas (2005a,b,c), is also an interesting extension. In practice, it is often the developed countries that are reluctant to have FTAs with less developed countries. In many cases, it is because they want to protect politically sensitive (import-competing) industries such as agriculture. We can broadly interpret our results to claim that developed countries are reluctant to have the FTAs since the political costs of opening such sensitive market is large and hence the direct surplus effect (including the political costs) is negative and large in magnitude. In order to address this issue more properly, however, we should explicitly reformulate the problem in the political economy framework.

We can also enrich the model by adding more industries with possibly different degrees of substitution within each sector. Extending the model to a dynamic setting with far-sighted governments is important, but is more challenging unless the number of countries is restricted to three or four.

Appendix

Proof of Lemma 2. It follows from (6) that

$$\begin{aligned}\sum_{k=1}^{n} s^k q_k^i(\mathbf{t}^i) &= \frac{1}{2-\sigma} - \frac{1}{2(1-\sigma)}\sum_{k=1}^{n} s^k t_k^i + \frac{\sigma}{2(1-\sigma)(2-\sigma)}\bar{t}^i \\ &= \frac{1}{2-\sigma}(1-\bar{t}^i).\end{aligned}$$

By substituting this result and (6) into $\partial V^i/\partial t_j^i$ in Lemma 1, we obtain

$$\frac{\partial V^i}{\partial t_j^i} = s^j\left[-\frac{1-\sigma}{(2-\sigma)^2} + \frac{\sigma^2}{4(1-\sigma)(2-\sigma)^2}\bar{t}^i - \frac{1}{4(1-\sigma)}t_j^i\right].$$

Let t(γ) denote the bilateral tariff reform schedule between countries *i* and j. This schedule satisfies $t_j^i(\gamma) = (1-\gamma)t^i$ and $t_i^j(\gamma)=(1-\gamma)t^j$ for $\gamma \in [0,1]$, and hence $t_j^i(0) = t^i$ and $t_j^i(1) = 0$, for example. All other tariff rates are kept unchanged, i.e., $t_k^i(\gamma) = t^i$ and $t_k^j(\gamma) = t^j$ for any $k \neq i, j$. Notice that $\bar{t}^i$ also changes in the course of tariff reform such that $\bar{t}^i(\gamma) = \sum_{k \notin Ci \cup \{j\}} s^k t^i + s^j(1-\gamma)t^i = (1 - s^{C_i} - \gamma s^k)\ t^i$, and similarly for $\bar{t}^j(\gamma)$. By substituting $\bar{t}^i(\gamma)$ and $t_j^i(\gamma)$ for $\bar{t}^i$ and t_j^i, respectively, and using $\mathrm{d}t_j^i/\mathrm{d}\gamma = -t^i$, we obtain

$$\frac{dV^i(\mathbf{t}^i(\gamma))}{d\gamma} = s^j t^i\left[\frac{1-\sigma}{(2-\sigma)^2} - \frac{\sigma^2}{4(1-\sigma)(2-\sigma)^2}(1-s^{C_i}-\gamma s^j)t^i + \frac{1}{4(1-\sigma)}(1-\gamma)t^i\right].$$

By integrating over γ, the welfare change of country *i* due to the FTA with *j* becomes

$$\begin{aligned}\Delta V^i(\mathbf{t}^i) &= s^j t^i\left[\frac{1-\sigma}{(2-\sigma)^2} - \frac{\sigma^2}{4(1-\sigma)(2-\sigma)^2}(1-s^{C_i}-\frac{s^j}{2})t^i + \frac{1}{8(1-\sigma)}t^i\right] \qquad (21)\\ &= \frac{s^j t^i}{8(1-\sigma)(2-\sigma)^2}\{8(1-\sigma)^2 + [4(1-\sigma) - (1-s^{C_i}-s^j)\sigma^2]t^i\}.\end{aligned}$$

The sufficient condition immediately follows. ■

Proof of Lemma 3. Recall the proof Lemma 2. The definition of the bilateral tariff reform schedule between countries *i* and j, denoted by $\mathbf{t}(\gamma)$, where $t_j^i(\gamma) = (1-\gamma)t$ and $\bar{t}^i(\gamma) = (1 - s^{C_i} - s^j\gamma)t = [1 - (|C_i|/n) - (\gamma/n)]\ t$, and similarly for *j*, while $\mathbf{t}^k(\gamma) = \mathbf{t}^k$ for any $k \neq i, j$, and any $\gamma \in [0,1]$. Then, it follows from (6) that

$$q_j^i(\mathbf{t}^i(\gamma)) = \frac{1}{2-\sigma} - \frac{1}{2(1-\sigma)}(1-\gamma)t + \frac{\sigma}{2(1-\sigma)(2-\sigma)}\left(1-\frac{|C_i|}{n}-\frac{\gamma}{n}\right)t,$$

$$q_k^i(\mathbf{t}^i(\gamma)) = \frac{1}{2-\sigma} - \frac{1}{2(1-\sigma)}t + \frac{\sigma}{2(1-\sigma)(2-\sigma)}\left(1 - \frac{|C_i|}{n} - \frac{\gamma}{n}\right)t.$$

Consequently, we have

$$\begin{aligned}
\frac{dq_j^i}{d\gamma} &= \frac{[n(2-\sigma)-\sigma]t}{2n(1-\sigma)(2-\sigma)}, \\
\frac{dq_k^i}{d\gamma} &= -\frac{\sigma\, t}{2n(1-\sigma)(2-\sigma)}.
\end{aligned}$$

Now, we can rewrite a change in country *i*'s industrial trade surplus.

$$\begin{aligned}
&\Delta[X^i(\mathbf{t}^{-i}) - M^i(\mathbf{t}^i)] \\
&= \int_0^1 \left[\frac{dX_j^i(\mathbf{t}^j(\gamma))}{d\gamma} - \frac{dM_j^i(\mathbf{t}^i(\gamma))}{d\gamma} - \sum_{k\neq i,j}\frac{dM_k^i(\mathbf{t}^i(\gamma))}{d\gamma}\right] d\gamma \\
&= \frac{1-\sigma}{n}\int_0^1 \left[2q_i^j(\mathbf{t}^j)\frac{dq_i^j}{d\gamma} - 2q_j^i(\mathbf{t}^i)\frac{dq_j^i}{d\gamma} - \sum_{k\neq i,j} 2q_k^i(\mathbf{t}^i)\frac{dq_k^i}{d\gamma}\right] d\gamma \\
&= \frac{2t(1-\sigma)}{n}\int_0^1 \left\{\frac{\sigma(|C_i|-|C_j|)}{2n(1-\sigma)(2-\sigma)}\,\frac{[n(2-\sigma)-\sigma]t}{2n(1-\sigma)(2-\sigma)}\right. \\
&\quad \left. + \left(\frac{\sigma}{2n(1-\sigma)(2-\sigma)}\right)\sum_{k\neq i,j}\left[\frac{1}{2-\sigma} - \frac{1}{2(1-\sigma)}t + \frac{\sigma}{2(1-\sigma)(2-\sigma)}\left(1-\frac{|C_i|}{n}-\frac{\gamma}{n}\right)t\right]\right\} d\gamma \\
&= \frac{\sigma t}{n^2(2-\sigma)}\int_0^1 \left\{\frac{(|C_i|-|C_j|)[n(2-\sigma)-\sigma]t}{2n(1-\sigma)(2-\sigma)}\right. \\
&\quad \left. + \frac{n-2}{2-\sigma} - \frac{n-2}{2(1-\sigma)}t + \frac{\sigma(n-2)}{2(1-\sigma)(2-\sigma)}\left(\frac{n-|C_i|-\gamma}{n}\right)t\right\} d\gamma.
\end{aligned}$$

The value of this formula decreases with $|C_j|$ since $n(2-\sigma)-\sigma > 0$. Whereas it increases with $|C_i|$ since

$$\frac{n(2-\sigma)-\sigma}{2n(1-\sigma)(2-\sigma)} - \frac{\sigma(n-2)}{2n(1-\sigma)(2-\sigma)} = \frac{2n(1-\sigma)+\sigma}{2n(1-\sigma)(2-\sigma)} > 0.$$

As for the impact on a consumer's gross utility, recall again the proof of Lemma 2. Let $s^k = 1/n$ for all $k = 1, \ldots, n$. Then, we find that

$$\Delta V^i(\mathbf{t}^i) = \frac{t}{8n(1-\sigma)(2-\sigma)^2}\left\{8(1-\sigma)^2 + \left[4(1-\sigma) - \left(1-\frac{2|C_i|+1}{n}\right)\sigma^2\right]t\right\}$$

also increases with $|C_i|$. ■

Proof of Proposition 2. Substituting $|C_i| = 1$ and $|C_j| = n - 1$ into the formulae obtained in the proof of Lemma 3, we have

$$\begin{aligned}
&\Delta[X^i(\mathbf{t}^{-i}) - M^i(\mathbf{t}^i)] \\
&\geq \frac{\sigma t}{n^2(2-\sigma)} \int_0^1 \left\{ \frac{(2-n)[n(2-\sigma)-\sigma]t}{2n(1-\sigma)(2-\sigma)} \right. \\
&\quad \left. + \frac{n-2}{2-\sigma} - \frac{n-2}{2(1-\sigma)} t + \frac{\sigma(n-2)(n-1-\gamma)t}{2n(1-\sigma)(2-\sigma)} \right\} d\gamma \\
&= \frac{(n-2)\sigma t}{2n^2(1-\sigma)(2-\sigma)^2} \int_0^1 \left\{ 2(1-\sigma) - 2(2-\sigma)t + \frac{(n-\gamma)\sigma t}{n} \right\} d\gamma \\
&= \frac{(n-2)\sigma t}{2n^2(1-\sigma)(2-\sigma)^2} \left[2(1-\sigma) - 2(2-\sigma)t + \frac{(2n-1)\sigma t}{2n} \right],
\end{aligned}$$

and

$$\Delta V^i(\mathbf{t}^i) \geq \frac{t}{8n(1-\sigma)(2-\sigma)^2} \left[8(1-\sigma)^2 + 4(1-\sigma)t - \frac{(n-3)\sigma^2 t}{n} \right].$$

Thus,

$$\begin{aligned}
\Delta u^i &= \Delta V^i + \Delta[X^i(\mathbf{t}^{-i}) - M^i(\mathbf{t}^i)] \\
&\geq \frac{t}{8n(1-\sigma)(2-\sigma)^2} \left[8(1-\sigma)^2 + 4(1-\sigma)t - \frac{(n-3)\sigma^2 t}{n} \right] \\
&+ \frac{(n-2)\sigma t}{2n^2(1-\sigma)(2-\sigma)^2} \left[2(1-\sigma) - 2(2-\sigma)t + \frac{(2n-1)\sigma t}{2n} \right] \\
&= \frac{t}{8n^2(1-\sigma)(2-\sigma)^2} [8(1-\sigma)(n-2\sigma) - A(\sigma, n)t],
\end{aligned}$$

where $A(\sigma, n) \equiv -4n + 4(5n-8)\sigma - [11n - 23 + (4/n)]\sigma^2$. It is now obvious that $\Delta u_i \geq 0$ if and only if either (i) $A(\sigma, n) \leq 0$ or (ii) $t \leq \tau(\sigma, n) \equiv 8(1-\sigma)(\mathrm{n} - 2\sigma)/A(\sigma, n)$ when $A(\sigma, n) > 0$ is satisfied.

Next , we show that $\tau(\sigma, n)$is decreasing in $\sigma \in (0,1)$ for any $n \geq 3$. We have

$$\begin{aligned}
\frac{\partial \tau}{\partial \sigma} &= -\frac{8}{A(\sigma,n)^2} \{ A(\sigma,n)[n - 2\sigma + 2(1-\sigma)] \\
&\quad + (1-\sigma)(n-2\sigma)\left[4(5n-8) - 2\sigma\left(11n - 23 + \frac{4}{n} \right) \right] \Bigg\} \\
&= -\frac{8}{A(\sigma,n)^2} \{ 2(1-\sigma)A(\sigma,n) \\
&\quad + (n-2\sigma)\left[4(5n-8) - 4n + \sigma(\sigma-2)\left(11n - 23 + \frac{4}{n} \right) \right] \Bigg\}.
\end{aligned}$$

Since $2(1-\sigma)\,A(\sigma, n) > 0$ and $n - 2\sigma > 0$, what remains to be shown is that the expression in the square brackets is positive. Now, $11n - 23 + (4/n) > 0$ for any $n \geq 3$ and $\sigma(\sigma - 2)$ takes its minimum of -1 at $\sigma = 1$, so that we have

$$\begin{aligned}
&4(5n-8)-4n+\sigma(\sigma-2)\left(11n-23+\frac{4}{n}\right)\\
&\geq 4(5n-8)-4n-\left(11n-23+\frac{4}{n}\right)\\
&=5n-9-\frac{4}{n},
\end{aligned}$$

which is positive for $n \geq 3$. ■

References

Bagwell, K. and R.W. Staiger (1997a), 'Multilateral tariff cooperation during the formation of customs unions', *Journal of International Economics* **42**, 91–123.

Bagwell, K. and R.W. Staiger (1997b), 'Multilateral tariff cooperation during the formation of free trade areas', *International Economic Review* **38**, 291–319.

Bagwell, K. and R.W. Staiger (1999a), 'An economic theory of GATT, *American Economic Review* **89**, 215–248.

Bagwell, K. and R.W. Staiger, (1999b), 'Regionalism and multilateral tariff co-operation', in J. Piggott and A. Woodland (eds.), *International Trade Policy and the Pacific Rim: Proceedings of the IEA Conference held in Sydney, Australia*, London: Macmillan, pp. 157–185.

Bagwell, K. and R.W. Staiger (2005), 'Multilateral trade negotiations, bilateral opportunism and the rules of GATT/WTO', *Journal of International Economics* **67**, 268–294.

Baldwin, R.E. (1995), 'A domino theory of regionalism', in R.E. Baldwin, P. Haaparanta and J. Kiander (eds.), *Expanding Membership of the European Union*, Cambridge: Cambridge University Press, pp. 25–48.

Bhagwati, J. (1993), 'Regionalism and multilateralism: an overview', in J. de Melo and A. Panagariya (eds.), *New Dimensions in Regional Integration*, Cambridge: World Bank and Cambridge University Press, pp. 22–51.

Bhagwati, J. and A. Panagariya (1996), 'Preferential trading areas and multilateralism–strangers, friends, or foes?', in J. Bhagwatiand and A. Panagariya (eds.), *The Economics of Preferential Trade Agreements*, Washington, DC: AEI Press, pp. 1–78.

Bloch, F. and M.O. Jackson (2004), 'The formation of networks with transfers among players', unpublished manuscript.

Bond, E.W., C. Syropoulos and L.A. Winters (2001), 'Deepening of regional integration and multilateral trade agreements', *Journal of International Economics* **53**, 335–361.

Dutta, B., S. Ghosal and D. Ray (2005), 'Farsighted network formation', *Journal of Economic Theory* **122**, 143–164.

Dutta, B. and S. Mutuswami (1997), 'Stable networks', *Journal of Economic Theory* **76**, 322–344.

Ethier, W.J. (1998), 'Regionalism in a multilateral world', *Journal of Political Economy* **106**, 1214–1245.

Freund, C.L. (2000a), ‘Different paths to free trade: the gains from regionalism’, *Quarterly Journal of Economics* **115**, 1317–1341.

Freund, C.L. (2000b), ‘Multilateralism and the endogenous formation of preferential trade agreements’, *Journal of International Economics* **52**, 359–376.

Freund, C.L. (2000c), ‘Spaghetti regionalism’, Board of Governors of the Federal Reserve System, International Finance Discussion Papers, No. 680.

Furusawa, T. and H. Konishi (2005), ‘Free trade networks with transfers’, *Japanese Economic Review* **56**, 144–164.

Furusawa, T. and H. Konishi (2004), ‘A welfare decomposition in quasi-linear economies’, *Economics Letters* **85**, 29–34.

Grossman, G.M. and E. Helpman (1995), ‘The politics of free-trade agreements’, *American Economic Review* **85**, 667–690.

Goyal, S. and S. Joshi (2006), ‘Bilateralism and free trade’, *International Economic Review* **47**, 749–778.

Jackson, M.O. and A. Wolinsky (1996), ‘A strategic model of social and economic networks’, *Journal of Economic Theory* **71**, 44–74.

Kemp, M.C. and H.Y. Wan Jr. (1976), ‘An elementary proposition concerning the formation of customs unions’, *Journal of International Economics* **6**, 95–97.

Kennan, J. and R.G. Riezman (1990), ‘Optimal tariff equilibria with customs unions’, *Canadian Journal of Economics* **23**, 70–83.

Kowalczyk, C. and R. J. Wonnacott, (1992), ‘Hubs and spokes, and free trade in the Americas’, NBER Working Paper No. 4198.

Krishna, P. (1998), ‘Regionalism and multilateralism: a political economy approach’, *Quarterly Journal of Economics* **113**, 227–251.

Krugman, P.R. (1991), ‘The move toward free trade zones’, in *Policy Implications of Trade and Currency Zones: A Symposium Sponsored by the Federal Reserve Bank of Kansas City*, Kansas City: Federal Reserve Bank of Kansas City, pp. 7-41.

Levy, P.I. (1997), ‘A political-economic analysis of free-trade agreements’, *American Economic Review* **87**, 506–519.

Mukunoki, H. and K. Tachi (2006), ‘Multilateralism and hub-and-spoke bilateralism’, *Review of International Economics* **14**, 658–674.

Ottaviano, G.I.P., T. Tabuchi and J.-F. Thisse (2002), ‘Agglomeration and trade: revisited’, *International Economic Review* **43**, 409–436.

Ohyama, M. (1972), ‘Trade and welfare in general equilibrium’, *Keio Economic Studies* **9**, 37–73.

Ornelas, E. (2005a), ‘Endogenous free trade agreements and the multilateral trading system’, *Journal of International Economics* **67**, 471–497.

Ornelas, E. (2005b), ‘Rent destruction and the political viability of free trade agreements’, *Quarterly Journal of Economics* **120**, 1475–1506.

Ornelas, E. (2005c), ‘Trade creating free trade areas and the undermining of multilateralism’, *European Economic Review* **49**, 1717–1735.

Panagariya, A. and P. Krishna (2002), ‘On necessarily welfare-enhancing free trade areas’, *Journal of International Economics* **57**, 353–367.

Richardson, M. (1993), ‘Endogenous protection and trade diversion’, *Journal of International Economics* **34**, 309–324.

Shubik, M. (1984), *A Game-Theoretic Approach to Political Economy*, Cambridge: MIT Press.

Viner, J. (1950), *The Customs Union Issue*, New York: Carnegie Endowment for International Peace.

Watts, A. (2001), 'A dynamic model of network formation', *Games and Economic Behavior* **34**, 331–341.

Yi, S.-S. (1996), 'Endogenous formation of customs unions under imperfect competition: open regionalism is good', *Journal of International Economics* **41**, 153–177.

Yi, S.-S. (2000), 'Free-trade areas and welfare: an equilibrium analysis', *Review of International Economics* **8**, 336–347.

Group Decision-Making in the Shadow of Disagreement

Kfir Eliaz, Debraj Ray and Ronny Razin

Paper presented at the CTN Workshop 2004. Reprinted from *Journal of Economic Theory* **132**(1), January 2007, 236–273, with permission from Elsevier.

A model of group decision-making is studied, in which one of two alternatives must be chosen. While agents differ over alternatives, everybody prefers agreement to disagreement. Our model is distinguished by three features: private information regarding valuations, differing intensities in preferences, and the option to declare neutrality to avoid disagreement. There is always an equilibrium in which the majority is more aggressive in pushing its alternative, thus enforcing their will via both numbers and voice. However, under general conditions an aggressive minority equilibrium inevitably makes an appearance, provided that the group is large enough. Such equilibria invariably display a 'tyranny of the minority': the increased aggression of the minority always outweighs their smaller number, leading to the minority outcome being implemented with larger probability than the majority alternative. We fully characterize the asymptotic behavior of this model as group size becomes large, and show that all equilibria must converge to one of three possible limit outcomes.

1. Introduction

Group decision-making is the process by which a collective of individuals attempt to reach a required level of consensus on a given issue. One can crudely divide this process into two important components: the deliberation among members of the group and the aggregation of individual opinions into a single group decision. Traditionally, the literature on political economy has focused on the second component by modelling group decision-making as voting games. More recently,

We thank Eddie Dekel, Wolfgang Pesendorfer, Muhamet Yildiz and two anonymous referees for useful comments. Support from the National Science Foundation under grant no. 0421852 is gratefully acknowledged.

several authors have examined group deliberation by studying its role in aggregating private information.[1]

In this paper we emphasize another important aspect of group deliberation: the role it plays in allowing group members to bargain over the final decision while avoiding disagreement.

For many group decisions, disagreement, or failure to reach a consensus, is costly for all members. There are numerous instances of such environments. A government may need to formulate a long-run response to terrorism: individuals may disagree – often vehemently – over the nature of an appropriate response, but everyone might agree that complete inaction is the worst of the options. Jury members in the process of deliberation may disagree on whether or not the defendent is guilty; however, in most cases they all prefer to reach an agreement than to drag the deliberations on endlessly. An investigative committee looking into the causes of a riot, or a political assassination, or a corruption scandal, may be under significant pressure to formulate *some* explanation, rather than simply say they don't know. Or citizens may need to agree on a constitution under the threat of civil war if such agreement cannot be reached.

When facing a threat of disagreement, groups usually try to avoid reaching this outcome by allowing its members, either formally or informally, to declare 'neutrality'; effectively, to suggest that they do not care strongly about either alternative and will support any outcome that may be more forcefully espoused by others with more intense preferences. For instance, think of an academic department that meets to make an offer to one of several candidates. Different faculty members may disagree over the ranking of the candidates. To be sure, some faculty members will feel more strongly about the choices than others. However, no member wants to see the slot taken away by the Dean because the department could not agree on an offer. Because faculty members may be uncertain as to the rankings and intensities of their colleagues, those faculty members who do not feel strongly about the issue will be less vocal and willing to 'go with the flow', while those who feel strongly about their favorite candidate will argue aggressively in her favor.

Likewise, in the jury example mentioned above, members may disagree over whether or not the defendent is guilty. Moreover, some jury members would have stronger feelings about the matter than others. However, in most cases, all would want to reach *some* unanimous decision rather than end up with a hung jury.[2] Consequently, those jurors who feel strongly towards conviction or acquital would

[1] See Gerardi and Yariv (2003), Austen-Smith and Federsen (2002) and Coughlan (2000)

[2] A case in point is the recent trial of Lee Malvo, the younger of the two men accused in the D.C. sniper case. According to the interviews conducted with some of the jury members who sat on that trial, the jury was split between conviction and acquital. Even though conviction could mean the death penalty for the accused, some of the jurors who opposed conviction remarked that they felt it was more important to reach a unanimous decision then end up with a hung jury (New York Times, Dec. 24, 2003).

be more vocal during deliberation, while those who feel less strongly on the issue might not oppose either side in order to facilitate an agreement.

A threat of disagreement has profound implications for group decision-making. Above all, preference intensities play a critical role: the decisions of individuals within the group are based not only on their *ordinal* ranking of the available alternatives, but also on how *strongly* they feel towards each one. With cardinal preferences central to our discourse, it is possible to address several important questions left unanswered in the literature. Do individuals, who favor an outcome which is less likely to be favored by the majority, fight more aggressively for their cause than individuals who hold the majority view? Can such aggression be strong enough so that the minority alternative is indeed implemented with greater probability than the outcome favored by the majority? Do higher levels of required consensus better protect the implementation of such minority outcomes? What is the likelihood that group deliberation will end in disagreement? To answer these and other related questions, we propose a simple and tractable model of group decision-making in the shadow of disagreement. We proceed as follows.

A group of n agents must make a joint choice from a set of two alternatives, A or B. Each agent must either announce an alternative – A or B – or she can declare ‘neutrality’, in that she agrees to be counted, in principle, for either side. Once this is accomplished, we tally declarations for each alternative, *including the number of neutral announcements*. If, for an alternative, the resulting total is no less than some exogenously given supermajority, we shall call that alternative *eligible*.

Because neutral announcements are allowed for and counted on both sides, all sorts of combinations are possible: exactly one alternative may be eligible, or neither, or both. If *exactly* one alternative is eligible, that alternative is implemented. If neither is eligible – which will happen if there is a fierce battle to protect one’s favorite alternative – then no alternative is picked: the outcome is disagreement. If both are eligible – as will typically be the case when there are a large number of neutrals – each alternative is equally likely to be implemented.

Our objective is to capture the basic strategic considerations common to several situations in which disagreement is costly. In this sense the model is sparse but inclusive: disagreement (or the threat of it) is at center stage, there is preference heterogeneity – in the ordinal sense of course, but in a cardinal sense as well, and there is the possibility of avoiding disagreement by means of capitulation. We therefore believe that by analyzing the equilibria of this model, we can gain important insights into a wide variety of situations.

Several specific features of the model deserve comment. First, while the language of a voting model is often used, we do not necessarily have voting in mind. The exogenously given supermajority may or may not amount to full consensus or unanimity, and in any case is to be interpreted as some preassigned degree of consensus or social norm that the group needs to achieve. For instance, in many informal situations, it may be considered socially undesirable to choose an option objected to by at least one person.

Second, relative to existing literature the option to remain neutral is a novel feature of our model. At the same time, it is a natural ingredient in the examples discussed above. We only add here that the neutrality option may be interpreted in several ways. One formal institution that is related is approval voting: members of the group submit an 'approval' or 'disapproval' for each alternative. A voter who approves both alternatives is effectively declaring neutrality. Or consider group debate that effectively proceeds like a war of attrition: members who drop out are in essence declaring neutrality. In addition, we have already discussed several examples in which neutrality is an informal yet central feature of the decision-making process. One could also imagine several quasi-formal mechanisms that help individuals to avoid disagreement by allowing their vote to be counted in a way that ensures a win to one of the alternatives. For example, one could delegate his ballot to an impartial arbitrator, who appreciates the anxiety of all concerned to avoid disagreement, and is therefore interested in implementing some outcome. In short, one could interpret the neutrality declaration as the reduced form of some unspecified procedure, which is used to help avoid unnecessary disagreements.

Third, in the model eligibility is a 'zero-one' characteristic: either an alternative is eligible or it is not. Any outcome that passes the test of garnering the support (either actively by declaring the alternative, or passively by declaring neutrality) of the required supermajority, is deemed socially fit – or eligible – to be implemented. There is no sense in which one alternative is 'more eligible' than another. Hence, if both alternatives are eligible, then both are on equal footing in terms of the social approval received. We therefore assume that the group implements each of the alternatives with equal probability.

To be sure, the particular tie-breaking rule used by a group may vary across different situations. In some situations, the group may vote again and again until only one outcome becomes eligible. In other situations, group members may bargain over which outcome to implement. There may also situations in which the group would simply choose the eligible outcome with the most votes. Or an arbitrator or committee chair may break ties. The advantage of our approach is that it greatly simplifies the analysis and allows us to provide a full characterization of the equilibria. Section 7.1 and 7.2 discusses some of the implications of assuming an alternative tie-breaking rule.

Finally, we are interested in the 'intensity' of preferences for one alternative over the other, and how this enters into the decision to be neutral, or to fight for one's favorite outcome. Specifically, we permit each person's valuations to be independent (and private) draws from a distribution, and allow quite generally for varying cardinal degrees of preference. A corollary of this formulation is that *others* are not quite sure of how strongly a particular individual might feel about an outcome and therefore about how that individual might behave. This is one way in which uncertainty enters the model.

Uncertainty plays an additional role, in that no one is sure how many people favor one given alternative over the other. We do suppose, however, that there is a common prior – represented by an independent probability p – that an individual will (ordinally) favor one alternative (call it A) over the other (call it B). Without loss of generality take $p \leq ½$ If, in fact, $p < ½$ one might say that it is commonly known that people of 'type A' are in a minority, or more precisely in a *stochastic* minority. We shall see that these two types of uncertainty are very important for the results we obtain.

We provide a full characterization of this model and study a number of extensions and variations. Our main results highlight the important implications of a threat of disagreement.

Cardinal preferences play a key role. In any equilibrium, each individual employs a cutoff rule: there will exist some critical relative intensity of preference (for one alternative over the other) such that the individual will announce her favorite outcome if intensities exceed this threshold, and neutrality otherwise. If a rule exhibits a lower cutoff, then an individual using that rule may be viewed as being more 'aggressive': she announces her own favorite outcome more easily, and risks disagreement with greater probability.

Equilibria in which an individual of the majority type uses a lower cutoff (and is therefore more aggressive) than her minority counterpart may be viewed as favoring the majority: we call them *majority equilibria*. Likewise, equilibria in which the minority type employs a lower cutoff will be called *minority equilibria*.

Using an obvious parallel from the Battle of the Sexes, there are always 'corner' equilibria in which one side is 'infinitely' aggressive – i.e., uses the lowest cutoff – while the other side is cowed into declaring full neutrality. But the resemblance ends there. In the model we study, a simple and weak robustness criterion reveals such equilibria to be particularly fragile. Section 4.2.2 introduces the refinement and shows how it removes corner equilibria in which one side invariably gives up.

Majority equilibria always exist. There always exists an equilibrium in which the majority uses a more aggressive cutoff than the minority (Proposition 1). This is an interesting manifestation of the 'tyranny of the majority'.[3] Not only are the majority greater in number (or at least stochastically so), they are also more vocal in expressing their opinion. In response – and fearing disagreement – the minority are more cowed towards neutrality. So in majority equilibrium, group outcomes are doubly shifted towards the majority view, once through numbers, and once through greater voice.

[3] It is possible that our use of this term constitutes a slight abuse of terminology, given that the phrase is typically invoked in the context of simple majority rule. We deal with supermajorities, so the term 'tyranny' (of either majority or minority) here is used in the sense of more strident use of *voice*.

Minority equilibria exist for large group sizes. Proposition 2 establishes the following result: if the required supermajority μ is not unanimity (i.e., $\mu < 1$), and if the size of the stochastic minority p exceeds $1 - \mu$, then for all sufficiently large population sizes, a minority equilibrium must exist.

How large is large? To be sure, the answer must depend on the specifics of the model, but our computations suggest that in reasonable cases, population sizes of 8–10 (certainly less than the size of a jury!) are enough for existence. We interpret this to mean that our existence result not only applies to large populations, but also to committees, juries, academic departments, cabinets and other groups which are numbered in the tens rather than in the hundreds.

From one point of view this result seems intuitive, yet from others it is remarkable. Intuitively, as population size increases, the two types of uncertainty that we described – uncertainty about type and uncertainty regarding valuation intensity – tend to diminish under the strength of the Law of Large Numbers. This would do no good if $p < 1 - \mu$, for then the minority would neither be able to win, nor would it be able to block the majority. [Indeed, Proposition 3 in Section 5.2 shows that if $p < 1 - \mu$, then for large population sizes a minority equilibrium cannot exist.] But if p exceeds $1 - \mu$, the minority acquires the 'credibility' to block the wishes of the majority, or at least does so when the population is large enough.

The existence of minority equilibria is not monotone in the consensus level. For two reasons, however, the above notion of 'credible blocking' does not form a complete explanation. First, a credible block is not tantamount to a credible *win*. Indeed, it is easy to see that as μ goes up, the minority find it easier to block but also harder to win. So the previous result must *not* be viewed as an assertion that the minority is 'better protected' by an increase in μ. Indeed, as an example in Section 5.1 makes clear, this is not true. [Nevertheless, insofar as existence is concerned, the fact that $p > 1 - \mu > 0$ guarantees existence of minority equilibrium for large population sizes.]

Second – and this extends further the line of argument in the previous paragraph – the case of unanimity ($\mu = 0$) is special. Proposition 4 shows that there are conditions (on the distribution of valuations) under which a minority equilibrium *never* exists, no matter how large the population size is. So blocking credibility alone does not translate into the existence of a minority equilibrium in the unanimity case. In short, any 'intuitive explanation' for Proposition 2 must also account for these observations.

The minority win more often in a minority equilibrium. Recall that in a majority equilibrium, the majority will have a greater chance of implementing its preferred outcome on two counts: greater voice, and greater number. Obviously, this synergy is reversed for the minority equilibrium: there, the minority have greater voice, yet they have smaller numbers. One might expect the net effect of these two forces to result in some ambiguity. The intriguing content of Proposition 5 is that in a minority equilibrium, the minority must always implement its favorite action with greater probability than the majority. Whenever a minority equilibrium exists, voice more than compensates for number.

Even in large groups, both sides may put up a fight. All equilibrium sequences must have limit points that are one of these three. Two of the outcomes may be viewed as 'limit minority equilibria'. One of them exhibits a zero cutoff for the minority, and the other exhibits a positive minority cutoff which is nevertheless lower than the majority cutoff. The third outcome is a 'limit majority equilibrium' in which the cutoff used by the majority is zero. The striking feature of these outcomes is that under some conditions, neither side gives up even if the opposition uses a zero cutoff! In particular, we establish the necessary and sufficient conditions for the existence of these interior cutoffs and describe exactly what they are.

Even as group size grows large, agreement is reached with uniformly positive probability. Given that both sides may put up a fight in relatively large groups, one might expect that for sufficiently high supermajority requirements disagreement will be endemic. However, for all non-unanimity rules, the probability of disagreement not only stays away from one, but actually converges to zero along any equilibrium sequences which converges to a limit outcome in which one side uses a zero cutoff. For those equilibria that converge to the remaining minority outcome, we show that the probability of disagreement is bounded away from one even as the population size goes to infinity.

Our results show that a 'shadow of disagreement' may effectively induce groups to make decisions that take into account their members' preference intensities. In particular, individuals who support an outcome that is less likely (ex-ante) to be favored by the majority, may still be able to implement that outcome if they feel sufficiently strongly about it. However, our paper also suggests that in group decision-making the outcomes tend to be invariably biased in one direction or another. In majority equilibrium this is obvious. But it is also true of minority equilibrium. This lends some support to a commonly-held view that group decision-making tends to have some degree of extremism built into the process.[4]

2. Related Literature

One central result in our paper is that minorities may fight more aggressively and win. Of course, the well-known Pareto-Olson thesis (see Pareto, 1927; and Olson, 1965) suggests that minorities might put up a stronger fight when voting is costly.

[4] The phenomenon of 'group polarization' has been extensively studied in the social psychology literature, most notably in Myers and Lamm (1976) and Lamm and Myers (1978). A more recent experimental study of this phenomenon is Cason and Mui (1997). In the political science and law literature, the potential impact of group polarization on court decisions has been studied by Sunstein (2000, 2002, 2003).

This intuition is confirmed in some complete-information models with private voting costs (see Araki and Börgers, 1996; and Haan and Kooreman, 2003), though in other variants with incomplete information (e.g., Palfrey and Rosenthal, 1983; Ledyard, 1984; Campbell, 1999; Krasa and Polborn, 2004; and Goeree and Grosser, 2005), the majority still wins at least as frequently as the minority even when the minority fights harder, assuming that preference intensities do not differ across groups.[5]

Our model also features a 'cost of voting': it is the expected loss caused by disagreement. But this cost is a *public* bad, and it cannot be shifted from one voter to another. (In addition, the magnitude of this cost is determined endogenously in equilibrium.)

An important feature of our model is that individuals base their decision on how strongly they prefer one alternative to another. This feature is shared with several papers that investigate different mechanisms in which intensity of preferences determine individual voting behavior. Vote-trading mechanisms, in which voters can trade their votes with one another, have been analyzed in Buchanan and Tullock (1962) and have more recently been revisited by Philipson and Snyder (1996) and Piketty (1999). Cumulative voting mechanisms in which each voter may allocate a fixed number of votes among a set of candidates has been analyzed as early as in Dodgson (1984) and more recently revisited by Gerber et. al (1998), Jackson and Sonnenschein (2007) and Hortala-Vallve (2004). In a related vein, Casella (2005) introduces a system of storable votes, in which voters can choose to store votes in order to use them in situations that they feel more strongly about.

These papers take a normative approach to group decision making in an attempt to design optimal procedures. Our approach is different. We take a positive approach and focus on existing institutions that rely on supermajority rules. We argue that a threat of disagreement may push individuals to base their decisions not only on their ordinal preferences, but also on their preference intensities. At the same time, we do not claim that the decision protocol we analyze – a supermajority rule coupled with a neutrality option and a threat of disagreement – necessarily leads to an efficient outcome (though mechanism design in our context would certainly be an interesting research project).

In particular, our analysis highlights the importance of consensus and the fear of gridlock as a mechanism through which intensities of preferences are translated into the decision making process. In this context, Ponsati and Sákovicz (1996) is also related to the present paper. Indeed, their model is more ambitious in that they explicitly attempt to study the dynamics of capitulation in an ambient environment similar to that studied here. This leads to a variant on the war of

[5] Certainly, if minorites are sufficiently more zealous in the espousal of their favorite issue, they may fight more aggressively and win more often, as Campbell (1999) also shows.

attrition, and their goal is to describe equilibria as differential equations for capitulation times, at which individuals cease to push their favorite alternative.

3. The Model

3.1 The group choice problem

A group of n agents must make a joint choice from a set of two alternatives, which we denote by A and B. The rules of choice are described as follows:

(1) Each agent must either name an alternative – A or B – or she can declare 'neutrality', in that she agrees to be counted, in principle, for either side.

(2) If the total number of votes for an alternative plus the number of neutral votes is no less than some exogenously given supermajority m ($> n/2$), then we shall call that alternative *eligible*.

(3) If no alternative is eligible, no alternative is chosen: a state D (for 'disagreement') is the outcome.

(4) If a single alternative is eligible, then that alternative is chosen.

(5) If *both* alternatives are eligible, A or B are chosen with equal probability.

Recall that our tie-breaking rule follows from our view of eligibility as a 'zero-one' characteristic: either an alternative is eligible or it is not, so that there is no sense in which one alternative is 'more eligible' than another. The point is simply this: if no alternative is blocked, it matters little whether one alternative gets more votes than another – the preassigned degree of consensus (or at least the lack of opposition) has been achieved for both alternatives. This is not to suggest, however, that other tie-breaking rules are not worth exploring. An obvious contender is one in which the option with the most votes wins in case both pass the supermajority requirement. We discuss the implication of using this alternative tie-breaking rule in Section 7.2.

3.2 Valuations

Normalizing the value of disagreement to zero, each individual will have valuations (v_A, v_B) over A and B. These valuations are random variables, and we assume they are private information. Use the notation (v, v'), where v is the valuation of the favorite outcome ($\max\{v_A, v_B\}$), and v' is the valuation of the remaining outcome ($\min\{ v_A, v_B\}$). An individual will be said to be of *type A* if $v = v(A)$, and of type B if $v = v_B$. [The case $v_A = v_B$ is unimportant as we will rule out mass points below.]

Our first restriction is

[A.1] Each individual prefers either outcome to disagreement. That is, $(v, v') \gg 0$ with probability one.

In Section 7.5 we remark on the consequences of dropping the assumption that disagreement is worse than either alternative.

In what follows we shall impose perfect symmetry across the two types *except* for the probability of being one type or the other, which we permit to depart from ½ [The whole idea, after all, is to study majorities and minorities.]

[A.2] A person is type A with (iid) probability $p \in (0, ½]$, and is type B otherwise. Regardless of specific type, however, (v, v') are chosen independently and identically across agents.

3.3 The Game

First, each player is (privately) informed of her valuation (v_A, v_B). Conditional on this information she decides to announce either A or B, or simply remain neutral and agree to be counted in any direction that facilitates agreement. Because an announcement of the less-favored alternative alone is weakly dominated by a neutral stance, we presume that each player either decides to announce her own type, or to be neutral.[6] The rules in Section 2.1 then determine expected payoffs.

4. Equilibrium

4.1 Cutoffs

We reiterate, for clarity in what follows, that when we say a player 'announces an outcome', we mean that *only* that alternative is named by the player; she has forsaken neutrality.

Consider a player of a particular type, with valuations (v, v'). Define $q \equiv n - m$. Notice that our player only has an effect on the outcome of the game – that is, she is pivotal – in the event that there are *exactly* q other players announcing her favorite outcome. For suppose there are more than q such announcements, say for A. Then B cannot be eligible, and whether or not A is eligible, our player's announcement cannot change this fact. So our player has no effect on the outcome. Likewise, if there are strictly less than q announcements of A, then B is eligible whether or not A is, and our player's vote (A or neutral) cannot change the status of the latter.

Now look at the pivotal events more closely. One case is when there are precisely q announcements in favor of A, and $q + 1$ or more announcements favoring B. In this case, by staying neutral our agent ensures that B is the only eligible outcome and is therefore chosen. By announcing A she guarantees that neither outcome is eligible, so disagreement ensues. In short, by switching her announcement from neutral to A, our agent creates a personal loss of v'.

[6] For a similar reason we need not include the possibility of abstention. Abstention (as opposed to neutrality) simply increases the probability of disagreement, which all players dislike by assumption.

In the second case, there are q announcements or less in favor of B. In this case, by going neutral our agent ensures that A and B are both eligible, so the outcome is an equiprobable choice of either A or B. On the other hand, by announcing A, our agent guarantees that A is the *only* eligible outcome. Therefore by switching in this instance from neutral to announcing A, our agent creates a personal gain of $v - (v + v')/2$.

To summarize, let P^+ denote the probability of the former pivotal event (q compatriots announcing A, $q + 1$ or more announcing B) and P^- the probability of the latter pivotal event (q compatriots announcing A, q or less announcing B). It must be emphasized that these probabilities are not exogenous. They depend on several factors, but most critically on the strategies followed by the other agents in the group. Very soon we shall look at this dependence more closely, but notice that even at this preliminary stage we can see that our agent must follow a *cutoff rule*. For announcing A is weakly preferred to neutrality if and only if

$$P^-[v - (v + v')/2] \geq P^+v'.$$

Define $u \equiv \frac{v-(v+v')/2}{v'}$ Note that (by [A.1]) u is a well-defined random variable. Then the condition above reduces to

$$P^-u \geq P^+, \tag{1}$$

which immediately shows that our agent will follow a cutoff rule using the variable u.

Notice that we include the extreme rules of always announcing neutrality (or always announcing one's favorite action) in the family of cutoff rules. [Simply think of u as a nonnegative extended real.] If a cutoff rule does not conform to one of these two extremes, we shall say that it is *interior*.

By [A.2], the variable u has the same distribution no matter which type we are referring to. We assume

[A.3] u is distributed according to the atomless cdf F, with strictly positive density f on $(0, \infty)$.

4.2 Symmetric equilibrium

In this paper, we study symmetric equilibria: those in which individuals of the same type employ identical cutoffs.

4.2.1 Symmetric cutoffs

Assume, then, that all A-types use the cutoff u_A and all B-types use the cutoff u_B. We can now construct the probability that a *randomly chosen* individual will announce A: she must be of type A, which happens with probability p, and she must want to announce A, which happens with probability $1 - \mathrm{F}(u_A)$. Therefore the overall probability of announcing A, which we denote by λ_A, is given by

$$\lambda_A \equiv p[1 - \mathrm{F}(u_A)].$$

Similarly, the probability that a randomly chosen individual will announce B is given by

$$\lambda_B \equiv (1 - p)[1 - \mathrm{F}(u_B)].$$

With this notation in hand, we can rewrite the cutoff rule (1) more explicitly. First, add P^- to both sides to get

$$P^-(1 + u) \geq P^+ + P^-.$$

Assuming that we are studying this inequality for a person of type A, the right-hand side is the probability that exactly q individuals announce A, while the term P^- on the left-hand side is the joint probability that exactly q individuals announce A *and* no more than q individuals announce B. With this in mind, we see that the cutoff u_A must solve the equation

$$\binom{n-1}{q}\lambda_A^q \sum_{k=0}^{q}\binom{n-1-q}{k}\lambda_B^k(1-\lambda_A-\lambda_B)^{n-1-q-k}(1+u_A)$$
$$=\binom{n-1}{q}\lambda_A^q(1-\lambda_A)^{n-1-q}. \tag{2}$$

Likewise, the cutoff u_B solves

$$\binom{n-1}{q}\lambda_B^q \sum_{k=0}^{q}\binom{n-1-q}{k}\lambda_A^k(1-\lambda_A-\lambda_B)^{n-1-q-k}(1+u_B)$$
$$=\binom{n-1}{q}\lambda_B^q(1-\lambda_B)^{n-1-q}. \tag{3}$$

We will sometimes refer to these cutoffs as ‘equilibrium responses’, to emphasize the fact that u_A embodies not just a ‘best response’ by an individual but is also an ‘equilibrium condition’ among individuals of the same type, given the cutoff used by the other type. The term ‘equilibrium response’ captures the hybrid nature of the group response.

4.2.2 A refinement for equilibrium responses

At this stage, an issue arises which we would do well to deal with immediately. It is that a symmetric cutoff of ∞ is always an equilibrium response for any type to any cutoff employed by the other type, provided that $q > 0$. This is easy enough to check: if no member in group A is prepared to declare A in any circumstance, then no individual in that group will find it in her interest to do so either. This is because (with $q > 0$) no such individual is ever pivotal.

Hence the 'full neutrality cutoff' $u = \infty$ is always an equilibrium response. But it is an unsatisfactory equilibrium response for the following reason. Fix a particular person, say of type A. Perturb the strategy of her compatriots from full neutrality to one in which they do announce A for a tiny range of very high u-values. Below, we demonstrate that this will make our person announce A for all but a bounded range of u-values, *where the bound on this range is independent of the perturbation to the compatriots*.

Before we show this, let us distill a formal requirement from the discussion above. Focus on the A-types with domain variable u. To handle infinite cutoffs, define the variable $w \equiv u/(1 + u)$; obviously, the cutoffs with respect to u translate directly into cutoffs with respect to w. In particular, full neutrality is just a cutoff of 1 in w-space. Now suppose that a (symmetric) cutoff w^* is an equilibrium response to some cutoff used by the other type. We will say that such a cutoff is *fragile* if there exists $\varepsilon > 0$ such that if w is the cutoff used instead of w^*, an individual member of the group will prefer to use a cutoff that is at least ε-far from w^*, *no matter how close* w is to w^*.

Observe that this criterion is much weaker than 'tatonnement style' refinements which would examine whether a response close to the putative equilibrium would lead to a sequence of 'myopic' best responses away from the original response. Our criterion raises a red flag only when there is a *discontinuous jump* from the original actions following an arbitrarily small perturbation – this is the significance of the requirement that ε is uniform in the perturbation. If our criterion is violated, the equilibrium response under scrutiny fails – in a strong sense – to be robust: the tiniest mistakes by others will drive an individual 'far away' from the prescribed action.

It turns out that this criterion eliminates – and *only* eliminates – those equilibrium responses exhibiting full neutrality.

Observation 1. *An equilibrium response is fragile if and only if it is infinite (in u-space, equivalently equal to 1 in w-space).*

Half this observation is obvious. Look at (2), which determines the cutoff u_A for a member of type A, as a function of λ_B (which is determined by the cutoff of the other type and so is fixed for the discussion) and of λ_A (which is determined by the cutoff employed by the A-compatriots). If the equilibrium response in question is finite, then $\lambda_A > 0$, and u_A is uniquely defined and moves continuously in λ_A, so that the question of fragility does not arise.

Indeed, in all the cases in which $\lambda_A > 0$, (2) reduces to the simpler form

$$\sum_{k=0}^{q} \binom{n-1-q}{k} \lambda_B^k (1-\lambda_A-\lambda_B)^{n-1-q-k}(1+u_A') = (1-\lambda_A)^{n-1-q}. \tag{4}$$

where we are denoting our individual's cutoff by u_A' as a reminder that we have not imposed the symmetry condition yet. Notice that this value of u_A' is uniformly bounded, say, by some number $M < \infty$ no matter what values λ_A and λ_B assume, even if λ_A approaches zero. This is the source of the fragility of full-neutrality: when $\lambda_A = 0$, so that all compatriots employ an infinite cutoff, then $u_A = \infty$ is *a* solution, but this cutoff jumps to no more than M as soon as there is any perturbation to a positive value of λ_a.

Intuitively, consider an individual of type A, and entertain a small perturbation in the fully neutral strategy of her compatriots: they now use a very large cutoff, but not an infinite one. Now, in the event that our agent is pivotal, it must be that her group is very large with high probability, because her compatriots are only participating to a tiny extent, and yet there are q participants in the pivotal case. This means that group A is likely to win (conditional on the pivotal event), and our individual will want to declare A for all but a uniformly bounded range of her u-values.

Note that in the special case of unanimity ($q = 0$), full neutrality is *never* an equilibrium response, so no refinements need to be invoked.

Finally, it should be noted that weak dominance is not enough to rule out full neutrality. To see this consider the profile in which both groups use a cutoff of zero and so are always voting their type. In this case, when a voter of type A is pivotal, he knows for sure that there are more than q declarations of B. Therefore, this voter has a strict incentive to claim neutrality. Note however, that the above profile is the only profile against which neutrality is a strict equilibrium response for *every* type.

4.2.3 Equilibrium conditions

In summary, then, the arguments of the previous section permit us to rewrite the equilibrium conditions (2) and (3) as follows:

$$\alpha(u_A, u_B) \equiv (1+u_A)\sum_{k=0}^{q} \binom{m-1}{k} \pi^k (1-\pi)^{m-1-k} = 1, \tag{5}$$

and

$$\beta(u_A, u_B) \equiv (1+u_B)\sum_{k=0}^{q} \binom{m-1}{k} \sigma^k (1-\sigma)^{m-1-k} = 1, \tag{6}$$

where $m = n - q$, $\pi \equiv \lambda_B/1 - \lambda_A$, and $\sigma \equiv \lambda_A/1 - \lambda_B$.

We dispose immediately of a simple subcase: the situation in which there is simple majority and n is odd, so that q precisely equals $(n - 1)/2$. The following result applies:

Observation 2. *If $q = (n - 1)/2$, there is a unique equilibrium which involves $u_A = u_B = 0$.*

To see why this must be true, consult (5) and (6). Notice that when $q = (n - 1)/2$, it must be that $m - 1 = n - q - 1 = q$. So an equilibrium response must equal zero no matter what the size of the other group's cutoff. In words, there is no cost to announcing one's favorite outcome in this case. Recall that the only conceivable cost to doing so is that disagreement might result, but in the pivotal case of concern to any player, there are q compatriots announcing the favorite outcome, which means there are no more than $n - 1 - q = q$ opposing announcements. So disagreement is not a possibility.

In the remainder of the paper, then, we concentrate on the case in which a genuine supermajority is called for:

[A.4] $q < (n - 1)/2$.

The following observations describe the structure of response functions in this situation. [A.1]–[A.4] hold throughout.

Observation 3. *A symmetric response u_i is uniquely defined for each u_j, and declines continuously as u_j increases, beginning at some positive finite value when $u_j = 0$, and falling to zero as $u_j \to \infty$.*

Observation 4. *Consider the point at which type A's response crosses the 45° line, or more formally, the value $\overline{u}$ at which $\alpha(\overline{u}, \overline{u}) = 1$. Then type B's equilibrium response cutoff to $\overline{u}$ is lower than $\overline{u}$, strictly so if $p < ½$.*

While the detailed computations that support these observations are relegated to the Appendix, a few points are to be noted. First, complete neutrality is not an equilibrium response (it is fragile) even when members of the other group are *always* announcing their favorite alternative. The argument for this is closely related to the remarks made in Section 4.2.2 and we shall not repeat them here. On the other hand, 'full aggression' – $u = 0$ – is *also* never an equilibrium response except in the limiting case as the other side tends to complete neutrality. These properties guarantee that every equilibrium (barring those excluded in Section 4.2.2) employs interior cutoffs.

Observation 4 requires some elaboration. It states that *at the point where the equilibrium response of Group A leaves both sides equally aggressive* (so that $u_A = u_B = \overline{u}$), group B's equilibrium response leads to greater aggression. The majority takes greater comfort from its greater number, and therefore are more secure about being aggressive. There is less scope for disagreement. However, note the emphasized qualification above. As we shall see later, it will turn out to be important.

Figure 1 provides a graphical representation. Each response function satisfies observation 3, and in addition observation 4 tells us that the response function for A lies above that for B at the 45° line. We have therefore established the following proposition.

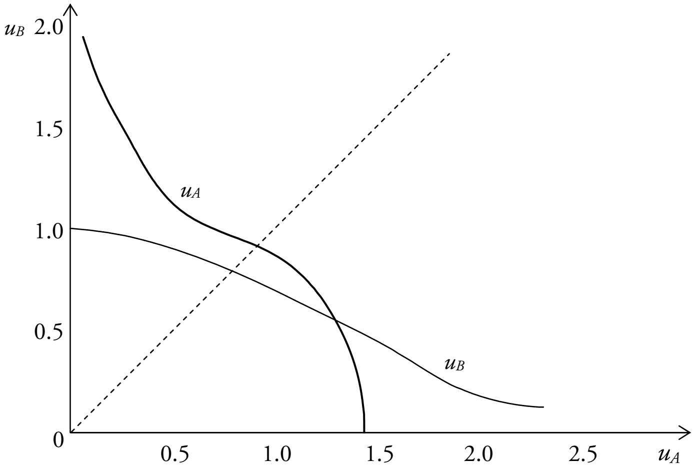


Figure1. Existence of a Majority Equilibrium

Proposition 1. *An equilibrium exists in which members of the stochastic majority – group B – behave more aggressively than their minority counterparts:* $u_B < u_A$.

Proposition 1 captures an interesting aspect of the 'tyranny of the majority'. Not only are the majority greater in number (at least stochastically so in this case), they are also more vocal in expressing their opinion. So group outcomes are doubly shifted – *in this particular equilibrium* – towards the majority view, once through numbers, and once through greater voice.[7] We will call such an equilibrium a *majority equilibrium*.

5. Minority Equilibria

5.1 Existence

Figure 1, which we used in establishing Proposition 1, is drawn from actual computation. We set $n = 4$, $p = 0.4$, $q = 1/4$, and chose F to be gamma with

[7] Notice that this model has no voting costs so that free-riding is not an issue. Such free-riding is at the heart of the famous Olson paradox (see Olson, 1965), in which small groups may be more effective than their larger counterparts.

parameters (3,4). Under this specification, there is, indeed, a unique equilibrium and (by Proposition 1) it must be the majority equilibrium.

Further experimentation with these parameters leads to an interesting outcome. When n is increased (along with q, to keep the ratio q/n constant), the response curves appear to 'bend back' and intersect yet again, this time above the 45° line (see Figure 2). A *minority equilibrium* (in which $u_A < u_B$, so that the minority are more aggressive) makes its appearance. For this example, it does so when there are 12 players.

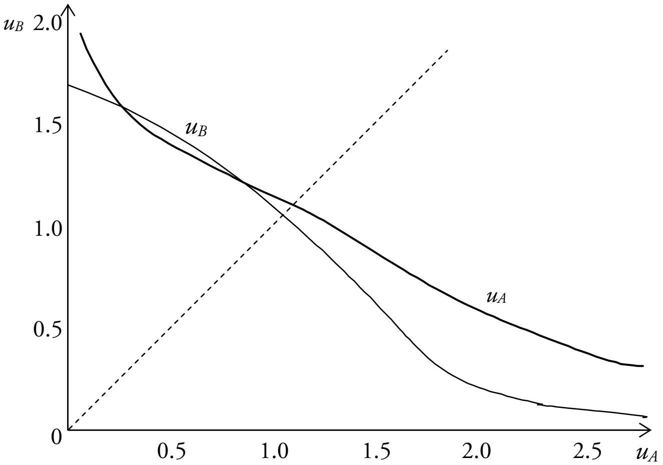


Figure 2. Minority Equilibrium

The bending-back of response curves to generate a minority equilibrium appeared endemic enough in the computations, that we decided to probe further. To do this, we study large populations in which the ratio of q to n is held fixed at $\nu \in (0, ½)$. More precisely, we look at sequences $\{n, q\}$ growing unboundedly large so that q is one of the (at most) two integers closest to νn. We obtain the following analytical confirmation of the simulations:

Proposition 2. *Assume that $0 < \nu < p \leq ½$. Consider any sequence $\{n, q\}$ such that $n \rightarrow \infty$ and q is one of the (at most) two integers closest to νn. Then there exists a finite N such that for all $n \geq N$, a minority equilibrium must exist.*

Several comments are in order. First, if there is a minority equilibrium, there must be at least two of them, because of the end point restrictions implied by Observations 3 and 4. Some of these equilibria will suffer from stability concerns similar to those discussed in Section 4.2.2. But there will always be other minority equilibria that are 'robust' in this sense.[8]

[8] Once again, this follows from the end-point restrictions.

Second, it might be felt that the threshold N described in Proposition 2 may be too large for 'reasonable' group sizes. Our simulations reveal that this is not true. For instance, within the exponential class of valuation distributions, the threshold at which a minority equilibrium appears is typically around $N = 10$ or thereabouts, which is by no means a large number.

Third, the qualification that $\nu > 0$ is important. The unanimity case, with $q = 0$ is delicate. We return to this issue in Section 7. The case $p \leq \nu$, which we also treat in next subsection, is of interest as well.

Finally, as an aside, note that Proposition 2 covers the symmetric case $p = ½$, in which case the content of the proposition is that an asymmetric equilibrium exists (for large n). To be sure, the proposition is far stronger than this assertion, which would only imply (by continuity) that a minority equilibrium exists (with large n) if p is sufficiently close to ½.

5.2 Discussion of the existence result

We can provide some intuition as to why minority existence is guaranteed for large n but not so for small n. Observe that when n is 'small', there are two sorts of uncertainties that plague any player. She does not know how many people there are of her type, and she is uncertain about the realized distribution of valuations. Both these uncertainties are troublesome in that they may precipitate costly disagreement. The possibility of disagreement is lowered by more and more people adopting a neutral stance, though after a point it will be lowered sufficiently so that it pays individuals to step in and announce their favorite outcome. For a member of the stochastic majority, this point will be reached earlier, and so a majority equilibrium will always exist.

On the other hand, when n is large, these uncertainties go away or at any rate are reduced. Now the expectation that the minority will be aggressive can be credibly self-fulfilling, because the expectation of an aggressive strategy can be more readily transformed into the expectation of a winning outcome. This intuition suggests that when the proportion of the minority is smaller than the superminority ratio, then minority equilibria do not exist for large n. This is confirmed in the following proposition.

Proposition 3. *Assume that $0 < p < \nu < ½$. Consider any sequence $\{n, q\}$ such that $n \to \infty$ and q is one of the (at most) two integers closest to νn. Then there exists a finite N such that for all $n \geq N$, a minority equilibrium does not exist.*

Taken together, Propositions 2 and 3 may suggest a monotonic relation between the supermajority requirement and the 'power' of the minority. Common intuition suggests that a higher supermajority requirement facilitates the emergence of a minority equilibrium. Indeed, the comparative politics literature compares different political systems and motivates what has been termed ‘consensus systems’ (Lijphart, 1999) by the desire to protect minorities from the tyranny of the majority.

However, this is generally false in our model. To see why, consider an individual of type A and her best response condition. As q decreases, A’s cutoff increases (holding B’s cutoff fixed), i.e., the group fights less aggressively. This follows from the fact that as q decreases, the probability that the B-types might block A increases. Because the above effect of lowering q applies to both groups, it is not clear which group benefits from this change.

To demonstrate the ambiguous effect of lowering q consider the following example: let $n = 1000$ (in light of Proposition 5 we intentionally pick a large n), $p = 0.4$ and consider the distribution function $F(u) = 1 - \frac{1}{\sqrt{\ln(u+e)}}$. For $q = 300$ there exists a minority equilibrium $u_A \simeq 1.35$ and $u_B \simeq 80$. However, for $q = 10$ there exists no minority equilibrium.

The above example seems to suggest that for some distribution functions a minority equilibrium may not exist when the supermajority requirement is at unanimity. Indeed, this is true.

Proposition 4. *Suppose that the distribution of u, F(u), satisfies the condition*

$$\frac{f(x)}{1-F(x)} \leq \frac{1}{(1+x)\ln(1+x)} \tag{7}$$

for all x > 0. Then in the case where m = n – i.e., unanimity – a minority equilibrium cannot exist for any n.

Note that cdf from the above example, $F(u) = 1 - \frac{1}{\sqrt{\ln(u+e)}}$, satisfies the sufficient condition (7). Moreover, while conceivably not necessary, *some* condition is needed to rule out minority equilibria in the unanimity case: there do exist cdf’s for which minority equilibria exist for all large n.[9]

Finally, compare and contrast our findings with the asymmetric equilibria in the Battle of the Sexes (BoS). Recall that analogues of those equilibria exist in this model as well, *but they have already been eliminated by the refinement introduced in Section 4.2.2*. One might suspect that the equilibria of our model converge (as n grows large) to the equilibria of the BoS game. In this sense, the equilibria could be perceived as purification of the BoS equilibria. However, Proposition 4 establishes that this is not the case. Indeed, in some cases, minority equilibria do not exist for any n. Hence, uncertainty plays a crucial role in our model. This conclusion will be further strengthened when we study limit outcomes in Section 6.

[9] One example of such a cdf is the exponential distribution $F(u) = 1 - e^{-u}$.

5.3 Minorities win in minority equilibrium

In this section we address the distinction between an equilibrium in which one group *behaves* more aggressively, and one in which that group *wins* more often. For instance, in the majority equilibrium the majority fights harder *and* wins more often than the minority does. [It cannot be otherwise, the majority are ahead both in numbers and aggression.] But there is no reason to believe that the same is true of the minority equilibrium. The minority may be more aggressive, but the numbers are not on their side.

However, a remarkable property of this model is that a minority equilibrium *must involve the minority winning with greater probability than the majority*. Provided that a minority equilibrium exists, aggression must compensate for numbers.

Proposition 5. *In a minority equilibrium, the minority outcome is implemented with greater probability than the majority outcome.*

This framework therefore indicates quite clearly how group behavior in a given situation may be swayed both by majority and minority concerns. When the latter occurs, it turns out that we have some kind of 'tyranny of the minority': they are so vocal that they actually swing outcomes (in expectation) to their side.

The proof of this proposition is so simple that we provide it in the main text, in the hope that it will serve as its own intuition.

Proof. Recall 5 and 6 and note that $u_A < u_B$ in a minority equilibrium. It follows right away that $\sum_{k=0}^{q}\binom{m-1}{k}\pi^k(1-\pi)^{m-1-k} > \sum_{k=0}^{q}\binom{m-1}{k}\sigma^k(1-\sigma)^{m-1-k}$, so that $\pi < \sigma$. Expanding this inequality, we conclude that $\lambda_B(1-\lambda_B) < \lambda_A(1-\lambda_A)$. Because $\lambda_A < ½$, this can only happen in two ways: either $\lambda_B > 1-\lambda_A$, or $\lambda_B < \lambda_A$. The former case is impossible, because λ_A and λ_B describe mutually exclusive events, so the latter case must obtain. But this implies the truth of the proposition. ■

6. Limit Equilibria

In Section 5.1 we established the existence of a minority equilibrium. Existence was guaranteed for large n and for all supermajority rules except for unanimity. As we've already remarked, there must be at least two such equilibria, while in addition we know that there is at least one majority equilibrium. This raises the question of what the set of equilibria look like as the group size grows without bound.

The purpose of this section is to prove that despite the possibly large multiplicity of equilibria for finite group size, there are exactly three limit outcomes. Two of these outcomes are 'limit minority equilibria'. Of the two, one exhibits a zero cutoff for the minority, and the other exhibits a positive minority cutoff which is nevertheless lower than the majority cutoff. The third outcome is a 'limit majority equilibrium' in which the cutoff used by the majority is zero.

Moreover, the two corner equilibria (in which one side always fights for its favorite) possess a special structure: *the other side does not necessarily yield fully*. That is, the rival side may use an interior cutoff even in the limit, and we will characterize this cutoff exactly.

We will also study disagreement probabilities along any sequence of equilibria.

6.1 A characterization of limit outcomes

We now study the various limit points of equilibrium cutoff sequences. We will denote a generic limit point by $(u_A{}^*, u_B{}^*)$.

Proposition 6. *Assume that* $v > 0$.

[1] *Suppose that* $(u^*_A, u^*_B) >> 0$. *Then both limits must be finite, and solve*

$$p[1 - \mathrm{F}(u^*_A)] = (1-p)[1-\mathrm{F}(u^*_B)] = v. \qquad (8)$$

[2] *Suppose that* $u^*_A = 0$. *Then* $u^*_B < \infty$ *if and only if* $p < (1 - v)/v$, *and in that case* u^*_B *is given by the condition*

$$F(u_B{}^*) = \frac{p(1-2v)}{(1-p)v}. \qquad (9)$$

[3] *Likewise, suppose that* $u_B{}^* = 0$. *Then* $u^*_A < \infty$ *if and only if* $1-p < (1-v)/v$, *and in that case* $u_A{}^*$ *is given by the condition*

$$F(u_A{}^*) = \frac{(1-p)(1-2v)}{pv}. \qquad (10)$$

[4] *Moreover, if* $p > v$, *each of the three configurations described above are limits for some sequence of equilibria.*

Proposition 6 is best understood by looking at Figure 3, which is drawn for the 'semi-corner case' in which $v < p < 1-p < v/(1-v)$. This figure depicts the loci $\lambda_B/(1-\lambda_A) = v/(1-v)$ and $\lambda_A/(1-\lambda_B) = v/(1-v)$, suitably truncated to respect the constraints that $\lambda_A \leq p$ and $\lambda_B \leq 1-p$. We claim that limit equilibrium cutoffs must simultaneously lie on *both* these truncated loci. To see this, suppose that some cutoff sequence $\{\lambda^n_A, \lambda^n_B\}$ lies below the locus $\lambda_B/(1-\lambda_A) = v/(1-v)$ (along some subsequence, but retain the original index n). Then the equilibrium condition (5) coupled with the strong law of large numbers, assures us that $u^n_A \to 0$, or that

$\lambda_A^n \to p$, which pulls the system back on to the locus. If, on the other hand, the cutoff sequence $\{\lambda_A^n, \lambda_B^n\}$ lies *above* the locus $\lambda_B / (1 - \lambda_A) = \nu / (1 - \nu)$, we have a contradiction as follows. First, by using (5) again, we may conclude that $\lambda_A^n \to 0$. Next, recall that $\lambda_B^n \leq 1 - p < \nu / (1 - \nu)$ (by assumption), but this and the previous sentence contradict the presumption that $\lambda_B^n / (1 - \lambda_A^n) > \nu / (1 - \nu)$ for all n.

Of course, the same sort of argument applies to both loci, so we may conclude that equilibrium cutoffs must converge to one of three intersections displayed in Figure 3.[10]

The last part of the proposition asserts that when minority equilibria exist for large n, each of the three cases indeed represent 'bonafide' limit points, in that each case is an attractor for some sequence of equilibria. For the majority corner, this is obvious, as majority equilibria always exist and no sequence of majority equilibria can ever converge to a minority outcome. That the other two limits are also non-vacuous follow from the proof of existence of minority equilibria (the reader is invited to study the formal arguments in Section 9).

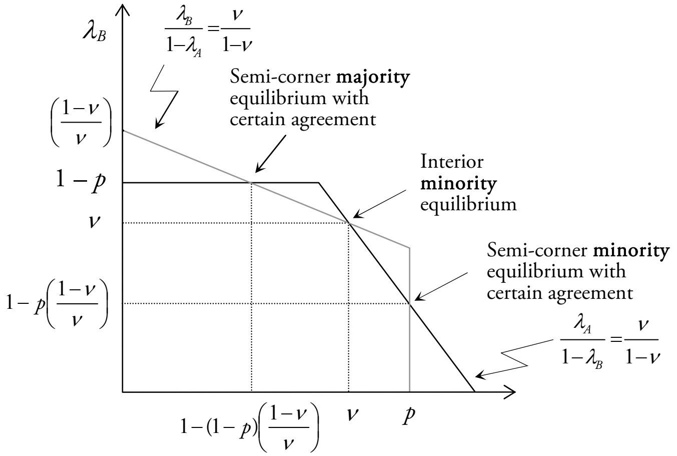

λ_B
$\frac{\lambda_B}{1-\lambda_A} = \frac{\nu}{1-\nu}$
Semi-corner **majority** equilibrium with certain agreement
$\left(\frac{1-\nu}{\nu}\right)$
$1-p$
ν
Interior **minority** equilibrium
Semi-corner **minority** equilibrium with certain agreement
$1-p\left(\frac{1-\nu}{\nu}\right)$
$\frac{\lambda_A}{1-\lambda_B} = \frac{\nu}{1-\nu}$
$1-(1-p)\left(\frac{1-\nu}{\nu}\right)$
ν
p

Figure 3. Limit equilibrium cutoffs

6.2 Disagreement

One important implication of Proposition 6 is that even when there is little uncertainty regarding the size of each faction, both sides may still put up a fight. In particular, when $1 - p < \frac{1-\nu}{\nu}$ all limit equilibria consist of 'fighting' on both sides. This raises the question of whether disagreement is bound to occur in large populations.

[10] It is also possible to construct versions of this diagram for the other cases, such as $1 - p > \nu/(1 - \nu)$ but $p < \nu/(1 - \nu)$

Proposition 7. *Assume $\nu > 0$.*

[1] *Suppose that $\nu < p < \frac{1-\nu}{\nu}$ and let u^*_B be the limit cutoff value that solves (9). Then in the limit semi-corner equilibrium $(0, u^*_B)$ both sides agree with certainty.*

[2] *Assume $1 - p < \frac{1-\nu}{\nu}$ and let u^*_A be the limit cutoff value that solves (10). Then in the limit semi-corner equilibrium $(u^*_A, 0)$ both sides agree with certainty.*

[3] *Consider any sequence of equilibria $(u^n_A, u^n_B) \rightarrow (u^*_A, u^*_B)$ where u^*_A and u^*_B solve (8). Then the probability of disagreement along that sequence is bounded away from one.*

The proofs of [1] and [2] follow immediately by looking at Figure 3. At the semi-corner minority equilibrium the proportion of *A* votes is simply *p*, which is strictly greater than ν. The proportion of *B* votes is $1 - p[(1 - \nu)/\nu]$, which is strictly smaller than ν. It follows that in the limit *A* is the unique eligible alternative, and hence that *A* will be implemented with certainty. Analogous arguments show that in the semi-corner majority equilibrium, *B* is the unique eligible alternative.

The proof of [3] is more involved. Recall that in this case the proportion of *A* and *B* votes converges to the superminority requirement ν. One may be tempted to conclude that the probability of disagreement in this case must converge to ¼. A closer examination reveals that this may not be the case. Indeed, what is important in determining the probability of disagreement is not the mere convergence of λ_A and λ_B to ν, but their *rate* of convergence. So far, the equilibrium conditions do not allow us to pin down the probability of disagreement in this case. Still, we establish that this probability is bounded away from one.

The intuition for this result is the following. Suppose that the probability of disagreement is high. Then the probability that each group is blocking the supermajority of its rival is also high. In particular, this means that group cutoffs are not wandering off to infinity. On the other hand, we can see that if group *A*, for example, is blocking group *B*, then the latter will be discouraged from making a *B* announcement. Doing so will most likely lead to disagreement, while casting a neutral vote ensures an agreement on *A*. This argument makes for high cutoffs, a contradiction to the bounded group cutoffs that were asserted earlier in this paragraph.

In part, the formalization of the above intuition is easy, but the simultaneous movements in population size and cutoffs necessitate a subtle argument. In particular, the last implication – that cutoffs become large with population size – rests on arguments regarding *rates* of change as a function of population. The reader is referred to the formal proof for details.

What allows individuals to agree, even when there are great many of them, is the option to remain neutral. This can be seen if we analyze a restricted version of our model in which individuals have only two options: *A* or *B*. We carry out this analysis in Section 7.3. There, we show that Proposition 7 ceases to hold.

Finally, note that the case of unanimity is *no*t covered here. This question remains open.

7. Extensions

7.1 Biased choice when both alternatives are eligible

Our model emphasizes majorities and minorities, but it can be used to study other issues. Consider the following example involving 'bias'. Suppose that an interested arbitrator or chair gets to implement the outcome in case both options are eligible.[11] To focus directly on the issue at hand, assume that the model is symmetric in every respect (inclusive of $p = ½$ though this is not logically needed for what follows) except for the bias, which we denote by $\alpha > ½$ in favor of alternative B.

It stands to reason that the presence of such a bias will spur A types on to greater aggression in pushing their alternative, while it might make the B types more complacent. This much is fairly obvious:[12] the question is whether such behavioral changes might nullify or even outweigh the bias.

The case of a strong bias, in which $\alpha \simeq 1$, is easiest to consider, because it has an unambiguous prediction:

***Observation 5**. Along any sequence of equilibria (as $\alpha \to 1$), it must be the case that $\lambda_B \to 0$, and $\lambda_A \to p = ½$.*

While a formal proof is postponed to Section 9, the intuition is simple. The B types know that as long as B is eligible, it is very likely to win. But pushing *just B* serves no additional purpose except to create a possible gridlock, which is damaging. Hence type B's equilibrium response must converge to 'full neutrality' as $\alpha \to 1$. For the A types, then, full aggression becomes an equilibrium response: they know that the eligibility of both alternatives is the same as an almost-sure defeat, and there is little likelihood of disagreement (given the timidity of the Bs).

The implication of these results is that the probability of A winning must converge to precisely the probability that the A types number more than q in the population. For A wins only when the A types block B, and triumph as the only eligible alternative. Otherwise it loses. If $q < n/2$ (so that we are dealing with supermajority rules), this probability must exceed ½. In contrast, when there is no bias, the model is completely symmetric and the probability that A wins must be no more than half, ex-ante.[13]

We have therefore shown that *arbitration biases against an alternative may increase the winning probability of that alternative, and indeed will increase it when the arbitration bias is infinitely high*.

[11] We owe this subsection to the comments of a referee.

[12] Formally, with multiple equilibria we would have to analyze changes in the equilibrium *correspondence*, but the reasonable conjecture in the main text can be easily made precise.

[13] The qualification 'no more than half' stems from the possibility of disagreement. However, remember that there may be multiple equilibria, so our statement in the text may be viewed as the outcome of symmetric randomization over all equilibria.

7.2 More on tie-breaking

The discussion in the previous section may be viewed more generally as an instance of various tie-breaking scenarios when both alternatives are eligible. For example, one might simply have a majority vote or some other 'runoff' in this case. The parameter α in Section 7.1 may be viewed as the reduced-form probability of win for type B in the runoff following eligibility of both alternatives. This makes little difference to the formalities of the model. One would simply redefine the variable u, depending on the value of α (the proof of Observation 5 in Section 9 does just this).

An interesting special case arises when α is given by a simple majority runoff. In this case, by Observation 2, α must equal $1-p$, a bias towards the majority. This be an additional source of minority aggression, as suggested by the analysis of the previous section.

Other tie-breaking procedures are harder to handle within our framework. For instance, suppose that the outcome with the more votes is chosen in the event that both outcomes are eligible. [The *existing* votes are recounted, so this is different from a runoff.] This leads to a more complicated setup; we indicate some of the steps.

Begin by deriving the necessary and sufficient condition for an individual of some type, say A, to weakly prefer an announcement of his favorite outcome – A in this case – to neutrality. To simplify the exposition we introduce the following notation. Define τ to be the joint probability that not counting our individual's vote, both A and B are eligible and both have the same number of declarations. Similarly, we define τ' to be the joint probability that not the A type's vote, both A and B are eligible, both have strictly less than q declarations, but B has exactly one declaration more than A. We also use the notation P^+ defined in Section 4.1.

Given the above tie-breaking rule, an A type weakly prefers to declare A than to declare neutrality if, and only if

$$\tau v + \tau'\left(\frac{v+v'}{2}\right) \geq P^+ v' + \tau' v' + \tau\left(\frac{v+v'}{2}\right)$$

Simplifying this inequality we obtain the following cutoff rule: declare A if, and only if

$$(\tau + \tau')u \geq P^+.$$

It follows that as in our original model, individuals base their decisions on how strongly they favor their preferred outcome to the alternative one. A similar inequality is obtained for the B types.

The complexity involved in analyzing our model under this alternative tie-breaking rule follows from the above inequality. Recall that in our original formulation the cutoff rule was expressed as the lower tail of a binomial distribution. Unfortunately, the new formulation does not accommodate such an expression.

Despite the added complexity, we are able to replicate some of our original results. First, it can be shown that all symmetric equilibria are interior (this is stated and proved as Observation 8 in Section 9). In contrast to the corresponding result in the paper (Observation 1), this result does not rely on any refinement. Second, a majority equilibrium always exists. This follows from arguments similar to those made in Proposition 1.

Establishing the existence of a minority equilibrium proved to be a formidable task. However, it is easy enough to generate numerical examples that exhibit the same features as those described in Proposition 2.[14]

7.3 No neutrality

In our opinion, when faced with impending disagreement, the option of a neutral stance is very natural. This is why we adopted this specification in our basic model. [As discussed already, neutrality is not to be literally interpreted as a formal announcement.] Nevertheless, it would be useful to see if the insights of the exercise are broadly preserved if announcements are restricted to be either A or B.

We can quickly sketch such a model. An individual is now pivotal under two circumstances. In the first event, the number of people announcing her favorite outcome is exactly q, which we assume to be less than $(n-1)/2$.[15] By announcing her favorite, then, disagreement is the outcome, while an announcement of the other alternative would lead to that alternative being implemented. The loss, then, from voting one's favorite in this event is precisely v' (recall that the disagreement payoff is normalized to zero). In the second event, the number of people announcing the alternative is exactly q. By announcing her favorite, she guarantees its implementation, while the other announcement would lead to disagreement. So the gain from voting one's favorite in this event is v. Consequently, an individual will announce her favorite if

$$\Pr(\text{exactly } q \text{ others vote for alternative})v \geq \Pr(\text{exactly } q \text{ others vote for favorite})v'.$$

Define $w \equiv v/v'$. Then equilibrium cutoffs w_A and w_B are given by the conditions

$$w_A \Pr(|B| = q) \geq \Pr(|A| = q) \tag{11}$$

and

[14] For example, a minority equilibrium exists for $F(u)=1-e^{-3u}$, $p = 0.4$, $n = 19$ and $q = 3$.

[15] The case $q = (n-1)/2$ is exactly the same as in Observation 2 for the main model. No matter what the valuations are, each individual will announce her favorite outcome.

$$w_B \Pr(|A| = q) \geq \Pr(|B| = q) \tag{12}$$

where $|A|$ and $|B|$ stand for the number of A- and B-announcements out of $n - 1$ individuals, and where equality must hold in each of the conditions provided the corresponding cutoff strictly exceeds 1, which is the lower bound for these variables.

In this variation of the model, it is obvious that at least one group must be 'fully aggressive' (i.e., its cutoff must equal one).[16] Moreover, as long as we are in the case $q < (n - 1)/2$, *both* groups cannot *simultaneously* be 'fully aggressive': one of the cutoffs must strictly exceed unity.

So, in contrast to our model, in which all (robust) equilibria are fully interior, the equilibria here are at 'corners' (full aggression on one side, full acquiescence on the other) or 'semi-corners' (full aggression on one side, interior cutoffs on the other). The semi-corner equilibria are always robust in the sense of Section 4.2.2, and we focus on these in what follows.[17]

In particular, to examine possible minority equilibria, set $w_A = 1$. Then use the equality version of (12) to assert that

$$w_B = \left(\frac{p + (1-p)H(w_B)}{(1-p)[1 - H(w_B)]} \right)^{n-1-2q} \tag{13}$$

in any such equilibrium, where H is the (assumed atomless) cdf of w, distributed on its full support $[1, \infty)$.

It is easy to use (13) to deduce

Observation 6.

[1] *A semi-corner minority equilibrium exists if (n, q) are sufficiently large.*

[2] *In any minority equilibrium, the minority outcome is implemented with greater probability than the majority outcome.*

So the broad contours of our model can be replicated in this special case. This is reassuring, because it reassures us of the robustness of the results. At the same time this variation allows us to highlight the main implication of allowing voters to remain neutral: absent neutrality voters may be locked into situations in which they are almost certain to disagree. This is formalized in the next result.

Observation 7. *Assume* $0 < \nu < p < ½$. *Consider any sequence* $\{n, q\}$ *such that* $n \to \infty$ *and* q *is one of the two integers closest to* νn. *Then there exists a sequence of semi-corner minority equilibria for which the probability of disagreement coverges to one.*

[16] Simply examine (11) and (12) and note that both right-hand sides cannot strictly exceed one.

[17] In contrast to our setup, the 'full corner' equilibria may or may not be robust. We omit the details of this discussion.

The above result demonstrates the importance of being neutral: neutrality allows the players to avoid disagreement. Recall that Proposition 7 establishes that with neutrality, the probability of disagreement at every interior equilibrium is bounded away from one. Once the option of neutrality is taken away, the probability that players reach a disagreement (at any interior equilibrium) must go to one along some sequence of minority equilibria.

7.4 Known group size

Our model as developed has the potential drawback that the instance of a known group size is not a special case. More generally, individuals may have substantial information regarding the ordinal stance of others (though still remaining unsure of their cardinal preferences).[18]

One way to accommodate this concern is to amend the model to posit a probability distribution $\theta(n_A)$ over the number n_A of A-types in the population. [The current specification of cardinal intensities may be retained.] This has the virtue of nesting our current model as well as known group size as special cases.[19] In addition, the basic structure of our model is easily recreated in this more general setting. For instance, if θ exhibits full support, a similar robustness argument applies to eliminate the 'coordination-failure' corner equilibria, and downward-sloping 'reaction functions', as in Figure 1, may be constructed just as before. The concept of a stochastic minority can also be easily extended. However, there are interesting conceptual issues involved in *changing* group size: in particular, we will need to specify carefully how θ alters in the process.

While a full analysis of this model is 'beyond the scope of the current paper', we provide some intuition by studying the extreme case in which group size is known; i.e., $\theta(n_A) = 1$ precisely at some integer $n_A < n/2$. We retain all our other assumptions.

Of course, θ no longer has full support, so the arguments in Section 4.2.2 do not apply to this case. To see why, consider the case when all B types are voting for B, whereas only extreme A-types are voting for A. When an A-type knows exactly how many B-types there are, he realizes that he can only create a disagreement by voting for A. Therefore, when group sizes are known, the two corner equilibria are robust (in the sense of Section 4.2.2). This suggests that the corner equilibria are unnatural in the following sense: when faced with some uncertainty about group sizes, some individuals may still put up a fight.

A further observation relates to the importance of group size in the emergence of minority equilibria. Potentially, the existence of minority equilibria in our original model may be due to two types of uncertainties that are relaxed in large groups. First, as the number of individuals in the group increases, voters

[18] In our current model, such 'substantial information' is only possible if p is close to either 0 or 1.

[19] In the current model, $\theta(n_A) = (n/(n_A))p^{n_A}(1 - \mathrm{p}^{n-n_A}$ for some $p \in (0, ½)$.

have a more accurate estimate of the proportion of their types in the group. Second, as the population increases, each individual has a better picture of the distribution of intensities among his compatriots.

What if group sizes are known? Then it can easily be shown that the equilibrium cutoff for one type depend only on the equilibrium cutoff of the other type. More precisely, an equilibrium (u_A, u_B) satisfies the following equations,

$$(1+u_A)\sum_{k=0}^{q}\binom{n_B}{k}(F(u_B))^{n_B-k}(1-F(u_B))^k = 1$$

$$(1+u_B)\sum_{k=0}^{q}\binom{n_A}{k}(F(u_A))^{n_A-k}(1-F(u_A))^k = 1$$

where $n_A < n_B$ are the number of individuals of type A and B respectively.

It is straightforward to construct examples in which there does not exist a minority equilibrium for small n_A and n_B. For instance, take $F(u) = 1 - \frac{1}{\sqrt{\ln(u+e)}}$, $n_A = 2$, $n_B = 3$ and $q = 1$. For these values there exists a unique interior majority equilibrium, $u_A \approx 250$ and $u_B.\approx 0.22$. However, using arguments similar to those employed in Propositions 2 and 4, one can show that for large n a minority equilibrium exists and the probability of disagreement is bounded away from one. By simple stochastic dominance arguments, it can be shown that in any minority equilbrium the minority wins more often.

We conclude that certainty regarding the numbers of A and B types is not sufficient to generate a minority equilibrium; even when the numbers of A and B types are known, we still need n to be sufficiently large for the minority to prevail.

7.5 Types who prefer disagreement to the rival alternative

Suppose there exist types who rank disagreement above their second best alternative. Clearly, voting for the preferred alternative is weakly dominant for these types. Hence, in any interior equilibrium these individuals would vote their type. In this sense, incorporating these voters into our model is equivalent to adding aggregate noise. We believe that if the proportion of such types is sufficiently low, all of our results continue to hold.

8. Summary

We study a model of group decision-making in which one of two alternatives must be chosen. While group members differ in their valuations of the alternatives, everybody prefers some alternative to disagreement.

We uncover a variant on the ‘tyranny of the majority’: there is always an equilibrium in which the majority is more aggressive in pushing its alternative, thus enforcing their will via both numbers and voice. However, under very general conditions an aggressive minority equilibrium inevitably makes an appearance, provided that the group is large enough. This equilibrium displays a 'tyranny of the minority': it is always true that the increased aggression of the minority more than compensates for smaller number, leading to the minority outcome being implemented with larger probability than the majority alternative.

These equilibria are not to be confused with 'corner' outcomes in which a simple failure of coordination allows any one group to be fully aggressive and another to be completely timid, without regard to group size. Indeed, one innovation of this paper is to show how such equilibria are entirely non-robust when confronted with varying intensities of valuations, and some amount of uncertainty regarding such valuations. In fact, as we emphasize in the paper, minority equilibria don't always exist: they don't exist, in general, for low population sizes and in the unanimity case they may not exist for *any* population size.

We also fully characterize limit outcomes as population size goes to infinity. We show that there are exactly three limit outcomes to which all equilibria must converge. Two of these outcomes are 'limit minority equilibria'. Of the two, one exhibits a zero cutoff for the minority, and the other exhibits a positive minority cutoff which is nevertheless lower than the majority cutoff. The third outcome is a 'limit majority equilibrium' in which the cutoff used by the majority is zero. The two corner equilibria which display full aggression on one side do not, in general, force complete timidity on the rival side. We provide a complete characterization by providing necessary and sufficient conditions for the interiority of such cutoffs and describing exactly their values.

Finally, we address the question of disagreement as group size grows large. We show that the probability of disagreement must converge to zero along all equilibrium sequences that converge to the semi-corners identified above. For those equilibria that converge to the remaining interior minority outcome, we show that the probability of disagreement is bounded away from one as the population size goes to infinity. The option to remain neutral is crucial in obtaining this result. Observation 7 in Section 7 considers an extension in which the neutrality option is removed, and proves that there is always a sequence of equilibria (in group size) along which the probability of disagreement must converge to one.

While we focus on the positive aspects of supermajority rules, our analysis suggests an approach from the viewpoint of mechanism design. Under supermajority rules, the fear of possible disagreement induces agents to base their actions on their *cardinal* preferences, rather than just on their ordinal ranking as in simple majority. Individuals who care a lot about the final outcome will indeed risk disagreement. Thus supermajority rules in the shadow of disagreement plays a possible role in eliciting intensities. However, there are caveats. First, disagreement is costly. It remains to be seen whether groups would obtain a net benefit by committing to the use of this costly option. Second, as our analysis shows, what determines agent behavior are *relative*, not absolute preference intensities over the different outcomes (see also Hortala-Vallve, 2004). This is an important (and complicated) enough question that deserves to be addressed in a separate paper.

9. Proofs

Proof of Observation 3. For concreteness, set $i = A$ and $j = B$. Fix any $u_B \in [0,\infty)$. Recall that

$$\pi = \frac{\lambda_B}{1-\lambda_A} = \frac{(1-p)[1-F(u_B)]}{1-p[1-F(u_A)]},$$

so that π is continuous in u_A, with $\pi \to 1 - F(u_B)$ as $u_A \to 0$, and $\pi \to (1-p)[1-F(u_B)]$ as $u_A \to \infty$. Consequently, recalling (5) and noting that $q < (n-1)/2$, we see that $\alpha(u_A, u_B)$ converges to a number strictly less than one as $u_A \to 0$, while it becomes unboundedly large as $u_A \to \infty$. By continuity, then, there exists some u_A such that $\alpha(u_A, u_B)$=1, establishing the existence of a cutoff.

To show uniqueness, it suffices to verify that α is strictly increasing in u_A. Because the expression $\sum_{k=0}^{q}\binom{m-1}{k}\pi^k(1-\pi)^{m-1-k}$ must be decreasing in π, it will suffice to show that π itself is declining in u_A, which is a matter of simple inspection.

To show that the response u_A strictly decreases in u_B, it will therefore be enough to establish that α is also increasing in u_B. Just as in the previous paragraph, we do this by showing that π is decreasing in u_B, which again is a matter of elementary inspection.

Finally, we observe that $u_A \downarrow 0$ as $u_B \uparrow \infty$. Note that along such a sequence, $\pi \to 0$ regardless of the behavior of u_A. Consequently, $\sum_{k=0}^{q}\binom{m-1}{k}\pi^k(1-\pi)^{m-1-k}$ converges to 1 as $u_B \uparrow \infty$. To maintain the equality (5), therefore, it must be the case that $u_A \downarrow 0$.

Of course, all these arguments hold if we switch A and B. ■

Proof of Observation 4. Let $\overline{u}$ be defined as in the statement of this Observation. Define $\overline{\lambda}_A \equiv p[1-F(\overline{u})]$ and $\overline{\lambda}_B \equiv (1-p)[1-F(\overline{u})]$. Then

$$(1+\overline{u})\sum_{k=0}^{q}\binom{m-1}{k}\overline{\pi}^k(1-\overline{\pi})^{m-1-k} = 1, \tag{14}$$

where $\overline{\pi} \equiv \overline{\lambda}_B/(1-\overline{\lambda}_A)$. Now recall that σ in (6) is defined by $\sigma = \frac{\lambda_A}{1-\lambda_B}$, so that if we consider the corresponding value σ defined by setting $u_A = u_B = \overline{u}$, we see that

$$\overline{\sigma} \le \overline{\pi} \text{ if and only if } \overline{\lambda}_A(1-\overline{\lambda}_A) \le \overline{\lambda}_B(1-\overline{\lambda}_B).$$

But $\lambda_A \leq ½$ (because $p \leq ½$), so that the second inequality above holds if and only if $\bar{\lambda}_A \leq \bar{\lambda}_B$, and this last condition follows simply from the fact that $p \leq ½$.

So we have established that $\bar{\sigma} \leq \bar{\pi}$. It follows that

$$\sum_{k=0}^{q} \binom{m-1}{k} \bar{\pi}^k (1-\bar{\pi})^{m-1-k} \leq \sum_{k=0}^{q} \binom{m-1}{k} \bar{\sigma}^k (1-\bar{\sigma})^{m-1-k}$$

and using this information in (14), we must conclude that

$$\beta(\bar{u},\bar{u}) = (1+\bar{u}) \sum_{k=0}^{q} \binom{m-1}{k} \bar{\sigma}^k (1-\bar{\sigma})^{m-1-k} \geq 1. \tag{15}$$

Recalling that β is increasing in its first argument (see proof of Observation 3), it follows from (15) that type *B*'s equilibrium response to $\bar{u}$ is no bigger than $\bar{u}$.

Finally, observe that all these arguments apply with strict inequality when $p < ½$. ■

Proof of Proposition 1. For each $u_B \geq 0$, define $\varphi(u_B)$ by composing equilibrium responses: $\varphi(u_B)$ is *B*'s equilibrium response to *A*'s equilibrium response to u_B. By Observation 3, we see that *A*'s equilibrium response is a positive, finite value when $u_B = 0$, and therefore so is *B*'s response to this response. Consequently, $\varphi(0)>0$. On the other hand, *A*'s equilibrium response is precisely $\bar{u}$ when $u_B = \bar{u}$, and by Observation 4 we must conclude that $\varphi(\bar{u}) < \bar{u}$. Because φ is continuous (Observation 3 again), there is $u_B^* \in (0, \bar{u})$ such that $\varphi(u_B^*) = u_B^*$. Let u_A^* be type *A*'s equilibrium response to u_B^**. Then it is obvious that (u_A^*, u_B^*) is an equilibrium. Because $u_B^* < \bar{u}$, we see from Observation 3 that $u_A^* > \bar{u}$. We have therefore found a majority equilibrium. ■

Proposition 2 and some subsequent arguments rely on the following lemma.

Lemma 1. *Consider any sequence $\{n, q\}$ such that $n \to \infty$ and q is one of the two integers closest to ν_n. For any u_A satisfying*

$$p[1-F(\bar{u}_A)] > \nu, \tag{16}$$

there exists a finite N such that for all $n \geq N$, $\hat{u}_B^n > u_B^n > \bar{u}_A$ where u_B^n solves (5) with $u_A = \bar{u}_A$, and $\hat{u}_B^n$ solves (6) with $u_A = \bar{u}_A$.

Proof. Consider any sequence $\{n, q\}$ as described in the statement of the lemma. Because $p > \nu$, there exists a range of positive cutoff values satisfying inequality (16). Consider any such value $\bar{u}_A$.and denote $\bar{\lambda}_A \equiv p[1-F(\bar{u}_A)]$. There exists a finite n^* such that for all $n \geq n^*$,

$$\overline{\lambda}_A > \frac{q}{n-1} \simeq \nu$$

Note that there is also an associated sequence $\{m\}$ defined by $m_n \equiv n - q$.[20]

We break the proof up into several steps.

Step 1. We claim that there exists an integer M such that for each $m \geq M$ there is $u_B^m < \infty$ that solves the following equation:

$$\sum_{k=0}^{q} \binom{m-1}{k} (\pi_m)^k (1-\pi_m)^{m-1-k} = \frac{1}{1+\overline{u}_A} \tag{17}$$

where

$$\pi_m \equiv \frac{\lambda_B^m}{1-\overline{\lambda}_A}$$

and

$$\lambda_B^m \equiv (1-p)[1-F(u_B^m)]\,.$$

We prove this claim. Note that for all $n \geq n^*$, $1-p \geq p > q/(n-1)$, so that

$$\overline{\pi} \equiv \frac{(1-p)(n-1)}{m-1} > \frac{q}{m-1} \simeq \frac{\nu}{1-\nu}$$

for all $n \geq n^*$. Consequently, by the Strong Law of Large Numbers (SLLN),

$$\sum_{k=0}^{q} \binom{m-1}{k} \overline{\pi}^k (1-\overline{\pi})^{m-1-k} \to 0$$

as m and q grow to infinity. It follows that there exists M such that for all $m \geq M$ (and associated q),

$$\sum_{k=0}^{q} \binom{m-1}{k} \overline{\pi}^k (1-\overline{\pi})^{m-1-k} < \frac{1}{1+\overline{u}_A}. \tag{18}$$

For such m, provisionally consider $u_B^m = 0$. Then

$$\frac{\lambda_B^m}{1-\overline{\lambda}_A} = \frac{1-p}{1-p[1-F(\overline{u}_A)]}\,,$$

and using this in (16), we conclude that

[20] While correct notation would demand that we denote this sequence by m_n, we shall use the index m for ease in writing.

$$\pi_m = \frac{\lambda_B^m}{1-\overline{\lambda}_A} = \frac{1-p}{1-p[1-F(\overline{u}_A)]} > \frac{(1-p)(n-1)}{m-1} = \overline{\pi}.$$

Combining this information with (18), we see that if $u_B^m = 0$, then

$$\sum_{k=0}^{q} \binom{m-1}{k} \pi_m^k (1-\pi_m)^{m-1-k} < \frac{1}{1+\overline{u}_A}. \tag{19}$$

Next, observe that if u_B^m is chosen very large, then λ_B^m and consequently π_m are both close to zero, so that $\sum_{k=0}^{q} \binom{m-1}{k} \pi_m^k (1-\pi_m)^{m-1-k}$ is close to unity. It follows that for such u_B^m,

$$\sum_{k=0}^{q} \binom{m-1}{k} \pi_m^k (1-\pi_m)^{m-1-k} > \frac{1}{1+\overline{u}_A}. \tag{20}$$

Combining (19) and (20) and noting that the LHS of (17) is continuous in u_B^m, it follows that for all $m \geq M$ there exists $0 < u_B^m < \infty$ such that the claim is true.

Step 2. One implication of (17) in Step 1 is the following assertion: as $(m, q) \to \infty$,

$$\pi_m \to \nu/(1-\nu) \in (0, 1), \text{ and in particular } u_B^m \text{ is bounded.} \tag{21}$$

To see why, note that $\frac{1}{1+\overline{u}_A} \in (0,1)$. Using (17) and SLLN, it must be that $\pi_m \to \nu/(1-\nu) \in (0, 1)$ as $(m, q) \to \infty$. Recalling the definition of π_m it follows right away that u_B^m must be bounded.

Step 3. Next, we claim there exists an integer M^* such that

$$\text{For all } m \geq M^*,\ u_B^m > \overline{u}_A. \tag{22}$$

To establish this claim, note first, using (16), that

$$p[1-F(\overline{u}_A)] > \frac{q}{n-1} = \frac{\frac{q}{m-1}}{1+\frac{q}{m-1}} \geq \frac{\frac{q}{m-1}}{\frac{1-p}{p}+\frac{q}{m-1}},$$

where the last inequality follows from the assumption that $p \in (0, \frac{1}{2}]$, so that $\frac{1-p}{p} \geq 1$. A simple rearrangement of this inequality shows that

$$\frac{(1-p)[1-F(\overline{u}_A)]}{1-p[1-F(\overline{u}_A)]} > \frac{q}{m-1} \simeq \frac{\nu}{1-\nu}. \tag{23}$$

Now suppose, contrary to the claim, that $u_B^m \leq \overline{u}_A$ along some subsequence of m. Then on that subsequence,

$$\pi_m = \frac{\lambda_B^m}{1-\overline{\lambda}_A} = \frac{(1-p)[1-F(u_B^m)]}{1-p[1-F(\overline{u}_A)]} \geq \frac{(1-p)[1-F(\overline{u}_A)]}{1-p[1-F(\overline{u}_A)]}. \tag{24}$$

Combining (23) and (24), we may conclude that along the subsequence of m for which $u_B^m \leq \overline{u}_A$,

$$\inf_m \pi_m > \frac{\nu}{1-\nu},$$

which contradicts (21) of Step 2.

To prepare for the next step, let $\hat{u}_B^m$ denote the equilibrium response of the B-types to $u_A = \overline{u}_A$. That is,

$$\frac{1}{1+\hat{u}_B^m} = \sum_{k=0}^{q} \binom{m-1}{k} \sigma_m^k (1-\sigma_m)^{m-1-k}, \tag{25}$$

where

$$\sigma_m \equiv \frac{\overline{\lambda}_A}{1-\hat{\lambda}_B^m}$$

and

$$\hat{\lambda}_B^m \equiv (1-p)[1-F(\hat{u}_B^m)].$$

Step 4. There is an integer M^{**} such that for all $m \geq M^{**}$, $\hat{u}_B^m > u_B^m$. To prove this claim, suppose on the contrary that $\hat{u}_B^m \leq u_B^m$ along some subsequence of m. [All references that follow are to this subsequence.] Then

$$\sigma_m = \frac{\overline{\lambda}_A}{1-\hat{\lambda}_B^m} = \frac{p[1-F(\overline{u}_A)]}{1-(1-p)[1-F(\hat{u}_B^m)]} \geq \frac{p[1-F(\overline{u}_A)]}{1-(1-p)[1-F(u_B^m)]} = \frac{\overline{\lambda}_A}{1-\lambda_B^m}. \tag{26}$$

Recall from (21), Step 2, that $\frac{\lambda_B^m}{1-\overline{\lambda}_A} \to \frac{\nu}{1-\nu}$. Therefore $\lambda_B^m \to \overline{\lambda}_B$, where $\overline{\lambda}_B \equiv \frac{\nu}{1-\nu}(1-\overline{\lambda}_A)$. Recall from (16) that $\overline{\lambda}_A > \nu$, so that $\overline{\lambda}_B < \nu$ and in particular $\overline{\lambda}_B < \overline{\lambda}_A$. Because $p \leq 1/2$, so is $\overline{\lambda}_A$, and these last assertions permit us to conclude that $\overline{\lambda}_A (1 - \overline{\lambda}_A) > \overline{\lambda}_B (1 - \overline{\lambda}_B)$, or equivalently, that

$$\frac{\overline{\lambda}_A}{1-\overline{\lambda}_B} > \frac{\overline{\lambda}_B}{1-\overline{\lambda}_A}.$$

Using this information in 26) and recalling that $\lambda_B^m \to \overline{\lambda}_B$, we may conclude that

$$\liminf_{m\to\infty} \sigma_m \geq \frac{\overline{\lambda}_{\mathrm{A}}}{1-\overline{\lambda}_{\mathrm{B}}} > \frac{\overline{\lambda}_{\mathrm{B}}}{1-\overline{\lambda}_{\mathrm{A}}} = \frac{\nu}{1-\nu},$$

where the last equality is from (21). It follows from (25) that $\hat{u}_B^m \to \infty$. But this contradicts our supposition that $\hat{u}_B^m \leq u_B^m$ (that along a subsequence) because the latter is bounded; see (21) of Step 2. ■

Proof of Proposition 2. Consider any sequence $\{n, q\}$ as described in the statement of the proposition. Choose some cutoff $\overline{u}_A$ that satisfies (16). By Lemma 1, there is an integer N such that for all $n \geq N$, $\hat{u}_B^n > u_B^n > \overline{u}_A$. Define, for each $n \geq N$ and each $u_A \in (0, \overline{u}_A]$, $\psi^n(u_A)$ as the *difference* between B's equilibrium response to u_A and the value of u_B to which u_A is an equilibrium response. By Lemma 1 and Observation 3, ψ^n is well-defined and continuous on this interval. Using Observation 3 yet again, it is easy to see that (for each n) $\psi^n(u_A) < 0$ for small values of u_A, while the statement of Lemma 1 assures us that $\psi^n(\overline{u}_A) > 0$. Therefore for each n, there is $\tilde{u}_A^n \in (0, \overline{u}_A)$ such that $\psi^n(\tilde{u}_A^n) = 0$. If we define $\tilde{u}_B^n$ to be the equilibrium response to $\tilde{u}_A^n$, it is trivial to see that $(\tilde{u}_A^n, \tilde{u}_B^n)$ constitutes an equilibrium.

Finally, note that

$$\tilde{u}_A^n < \overline{u}_A < u_B^n < \hat{u}_B^n < \tilde{u}_B^n,$$

where the second and third inequalities are a consequence of Lemma 1, and the last inequality comes from the fact that the equilibrium response function is decreasing (Observation 2). This means that $(\tilde{u}_A^n, \tilde{u}_B^n)$ is a minority equilibrium. ■

Proof of Proposition 3. Suppose on the contrary that a minority equilibrium (u_A^n, u_B^n) exists along some subsequence of n (all references that follow are to this subsequence). Then $\lim_{n\to\infty}(u_A^n, u_B^n)$ is either (∞,∞), $(0,\infty)$ or a pair of strictly positive but finite numbers (u_A^*, u_B^*). To prove that our supposition is wrong, we show that none of these limits can apply.

Assume $(u_A^n, u_B^n) \to (\infty,\infty)$. Then $\lambda_A^n \to 0$ and $\lambda_B^n \to 0$. This implies that $\pi^n \to 0$ and $\sigma^n \to 0$. But this implies, by equations (5) and (6) and using SLLN, that $(u_A^n, u_B^n) \to (0,0)$, a contradiction.

Assume $(u_A^n, u_B^n) \to (0,\infty)$. Then $\lambda_A^n \to p$ and $\lambda_B^n \to 0$, so that $\sigma^n \to p < \nu < \frac{q}{m-1}$. But using (6) and SLLN, this implies that $u_B^n \to 0$, a contradiction.

Assume $(u_A^n, u_B^n) \to (u_A^*, u_B^*)$, where both u_A^* and u_B^* are strictly positive and finite. Using SLLN and equations (5) and (6), it follows that π^n and σ^n must both converge to $\frac{q}{m-1}$. This means that $\lambda_A^n \to \lambda_A^*$ and $\lambda_B^n \to \lambda_B^*$ such that

$$\frac{\lambda_B^*}{1-\lambda_A^*} = \frac{\lambda_A^*}{1-\lambda_B^*}.$$

This equality holds only if $\lambda_A^* = \lambda_B^*$, or if $\lambda_A^* = 1 - \lambda_B^*$. Suppose the former is true. Then $\pi^n \rightarrow \pi^*$ where

$$\pi^* = \frac{\lambda_B^*}{1-\lambda_A^*} < \frac{v}{1-v} \simeq \frac{q}{m-1}$$

But the above inequality implies, by (5) and SLLN, that $u_A^n \rightarrow 0,$ a contradiction. Suppose next that $\lambda_A^* = 1 - \lambda_B^*$. But $1 - \lambda_B^* > p > \lambda_A^*$, a contradiction. ■

*Proof of Proposition 4***.** Under unanimity, (5) and (6) reduce to

$$\frac{1}{1+u_A} = (1-\pi)^{n-1} \tag{27}$$

and

$$\frac{1}{1+u_B} = (1-\sigma)^{n-1} \tag{28}$$

For any given n and $k = A, B$, define $y_k \equiv (1+ u_k)^{1/(n-1)}$. Then $y_k \geq 1$, and (27) and (28) may be rewritten as

$$1-\pi = \frac{1}{y_A} \tag{29}$$

and

$$1-\sigma = \frac{1}{y_B}. \tag{30}$$

Recalling that $\pi = \lambda_B/(1 - \lambda_A)$ and $\sigma = \lambda_A/(1 - \lambda_B)$, we may use (29) and (30) to solve explicitly for λ_A and λ_B. Doing so and writing out λ_k for $k = A, B$, we see that

$$\lambda_A = p[1-F(u_A)] = \frac{y_B - 1}{y_A + y_B - 1}, \tag{31}$$

while

$$\lambda_B = (1-p)[1-F(u_B)] = \frac{y_A - 1}{y_A + y_B - 1}. \tag{32}$$

By multiplying both sides of (31) by $1 - F(u_B)$ and both sides of (32) by $1 - F(u_A)$ and using the fact that $p < 1 - p$, we may conclude that

$$[1 - F(u_B)][(1 + u_B)^{1/(n-1)} - 1] < [1 - F(u_A)][(1 + u_A)^{1/(n-1)} - 1] \tag{33}$$

We will now prove that $u_A > u_B$. Given (33), it will suffice to prove that

$$[1 - F(x)][(1+ x)^{1/(n-1)} - 1]$$

is nondecreasing in x. This, in turn, is implied by the stronger observation that

$$\frac{d}{dx}[1 - F(x)][(1+ x)^{1/(n-1)} - 1] \geq 0$$

for every $x > 0$, or equivalently, that

$$\frac{f(x)}{1-F(x)} \leq \frac{\theta(1+x)^{\theta-1}}{(1+x)^{\theta}-1}, \tag{34}$$

where $\theta \equiv \frac{1}{n-1} \in (0,1]$.

To this end, we demonstrate that for all $x > 0$ and $\theta \in (0, 1]$,

$$\frac{\theta(1+x)^{\theta-1}}{(1+x)^{\theta}-1} \geq \frac{1}{(1+x)\ln(1+x)}. \tag{35}$$

To establish (35), note that for fixed $x > 0$, $h(\theta) \equiv (1 + x)^{\theta}$ is differentiable and convex in x. By a standard property of differententiable convex functions, $h(\theta_1) - h(\theta_2) \leq h'(\theta_1)\ (\theta_1 - \theta_2)$ for all θ_1 and θ_2. Applying this inequality to the case $\theta_1 = \theta$ and $\theta_2 = 0$, we may conclude that

$$h(\theta) - h(0) = (1 + x)^{\theta} - 1 \leq h'(\theta)\theta = (1 + x)^{\theta}\ln(1 + x)\theta,$$

and a quick rearrangement of this inequality produces (35).

To complete the proof, combine (7) and (35) to obtain (34). ■

*Proof of Proposition 6***.** Recall the conditions describing equilibrium cutoffs:

$$\frac{1}{(1+u_A)} = \sum_{k=0}^{q}\binom{m-1}{k}\pi^k(1-\pi)^{m-1-k}$$

and

$$\frac{1}{(1+u_B)} = \sum_{k=0}^{q}\binom{m-1}{k}\sigma^k(1-\sigma)^{m-1-k}.$$

For each integer n (with associated m and q) and every $u \geq 0$, define a function $h(u, n)$ by the condition that

$$\sum_{k=0}^{q} \binom{m-1}{k} h(u,n)^k (1-h(u,n))^{m-1-k} \equiv \frac{1}{1+u}.$$

Note that h is well-defined for each (u, n). With this in hand, we may rewrite the equilibrium conditions more succintly as

$$\frac{\lambda_B^n}{1-\lambda_A^n} = \pi^n = h(u_A^n, n) \equiv \alpha^n \tag{36}$$

and

$$\frac{\lambda_A^n}{1-\lambda_B^n} = \sigma^n = h(u_B^n, n) \equiv \beta^n \tag{37}$$

where we are now starting to index all endogenous variables by n in order to prepare for sequences of equilibria. Solving these two equations for λ_A^n and λ_B^n, we see that

$$\lambda_A^n = p[1-F(u_A^n)] = \frac{\beta^n(1-\alpha^n)}{1-\alpha^n\beta^n} \tag{38}$$

and

$$\lambda_B^n = (1-p)[1-F(u_B^n)] = \frac{\alpha^n(1-\beta^n)}{1-\alpha^n\beta^n}. \tag{39}$$

We now study various limits of equilibrium cutoff sequences. We will denote the limits in all cases by (u_A^*, u_A^*). The following lemma summarizes simple properties of h and will be used throughout.

Lemma 2.
[1] *For every n, h is strictly increasing in u, with* $h(0, n) = 0$ *and* $h(u, n) \rightarrow 1$ *as* $n \rightarrow \infty$.
[2] *If* u^n *converges to u with* $0 < u < \infty$, *then* $\lim_{n\rightarrow\infty} h(u^n,n) = \nu/(1-\nu)$.
[3] *If* u^n *converges to 0 then* $\limsup_{n\rightarrow\infty} h(u^n,n) \leq \nu/(1-\nu)$.
[4] *If* $u^n \rightarrow \infty$, *then* $\liminf_{n\rightarrow\infty} h(u^n,n) \geq \nu/(1-\nu)$.

The proof of this lemma follows from routine computations and the use of the law of large numbers, and is omitted.

Now we prove part [1] of the proposition. First, we claim that u_A^* and u_B^* are finite. For suppose, say, that $u_A^* = \infty$ (the argument in the other case is identical). It follows from (38) that either α^n has a limit point at 1, or that β^n has a zero limit point. The latter possibility is ruled out by Lemma 2, because $u_B^* > 0$ by assumption. It follows that $\limsup_{n\to\infty}\alpha^n = 1$, but then Lemma 2 assures us that $u_B^* = \infty$ as well.

The first of the two conclusions in the preceding sentence implies that $\limsup_{n\to\infty}\lambda_B^n = 1$ (use (37)), but the second conclusion implies that $\lim_{n\to\infty}\lambda_B^n = 0$ (use (39)). These two implications contradict each other.

So $0 \ll (u_A^*, u_B^*) \ll \infty$, but we know then from Lemma 2 that $(\alpha^n,\beta^n) \to (\nu,\nu)$ as $n \to \infty$. Simple computation using (38) and (39) then yields (8). It should be noted that this limit (which is unique in the class of strictly positive limits) has $u_A^* < u_B^*$; that is, it is a 'limit' minority equilibrium.

Next, we prove part [2]; the proof of part [3] is completely analogous. Suppose, then, that $u_A^* = 0$. We first prove the sufficiency of the restriction on p. To this end, assume that $u_B^* =\infty$. Consider some subsequence in which α^n and β^n converge (to some α^* and β^*). Then (38) implies that

$$\frac{\beta^*(1-\alpha^*)}{1-\alpha^*\beta^*} = p \tag{40}$$

while at the same time, (39) implies that

$$\frac{\alpha^*(1-\beta^*)}{1-\alpha^*\beta^*} = 0, \tag{41}$$

(41) implies either that $\alpha^* = 0$ or that $\beta^*= 1$. But the latter cannot happen, for then (40) cannot be satisfied (note that the LHS of (40) is well-defined even when $\beta^* = 1$, because $\alpha^* < 1$ by Lemma 2). So it must be that $\alpha^* = 0$. But then (40) implies that $p = \beta^*$. Lemma 2 tells us that $\beta^* \geq \nu/(1 - \nu)$, so that $p \geq \nu/(1 - \nu)$.

Conversely, suppose that $u_B^* < \infty$. Again, consider some subsequence in which α^n and β^n converge to some α^* and β^*. Therefore (38) implies that

$$\frac{\beta^*(1-\alpha^*)}{1-\alpha^*\beta^*} = p \tag{42}$$

while (39) implies that

$$\frac{\alpha^*(1-\beta^*)}{1-\alpha^*\beta^*} = (1-p)\,[1 - F(u_B^*)] \tag{43}$$

We can eliminate α^* from this system. We also note that by Lemma 2, β^* must equal $\nu/(1 - \nu)$. Using these observations along with some routine computation, we obtain precisely (9).

We also know that $F(u_B^*) < 1$. Using this information in (9), we may conclude that $p < \nu/(1 - \nu)$.

Finally, we establish part [4]. Assume, to the contrary, there exists no sequence of equilibria whose limit is given by the first configuration. By parts [2] and [3] of the proposition, the limit of any sequence of minority equilibria has either $u_A^* = 0$ or $u_A^* > u_B^*$. To reach a contradiction, pick any $u_A > 0$ satisfying (16). By Lemma 1, there exists an integer N such that for all $n \geq N$, there exists a minority equilibrium $(u_A{}^n, u_B{}^n)$ with $u_A{}^n > u_A$. From Proposition 1 it follows that for any $p < (1/2)$ and for any n, there does not exist a pair of numbers (u, u) that solve the equilibrium conditions (5) and (6). We therefore conclude that for all $n \geq N$, there exists a minority equilibrium $(u_A{}^n, u_B{}^n)$ with $0 < u_A < u_A{}^n < u_B{}^n$, in contradiction to our initial assumption.

Suppose next that there exists no sequence of equilibria whose limit is given by the second configuration. Then by parts [1] and [3] of the proposition, the limit of any minority equilibrium must satisfy that $u_A^* \geq \nu > 0$. Let $\varepsilon \in (0, \nu)$. By Lemma 1, there exists a finite $N > 0$ such that for all $n \geq N$ there exists a minority equilibrium $(u_A{}^n, u_B{}^n)$ with $u_A{}^n < \varepsilon$. But this means that the limit of any such sequence cannot satisfy that $u_A^* \geq \nu$, a contradiction.

Finally, assume there exists no sequence of equilibria whose limit is given by the third configuration. This implies, by [1] and [2], that the limit of any sequence of equilibrium cutoffs has $u_A^* < u_B^*$. But this contradicts Proposition 1, which states that for every n there exists an equilibrium with $u_A{}^n > u_B{}^n$. ■

Proof of Proposition 7. The proofs of [1] and [2] are given in the discussion following the statement of the proposition in the text. We now proceed to prove [3]. Assume that $q < \frac{n-1}{2}$ (When $q = \frac{n-1}{2}$ the probability of disagreement is zero). Note that the probability of disagreement is equal to $\Pr(|A| > q, |B| > q)$, where $|\cdot|$ stands for cardinality. Because

$$\Pr(|A| > q, |B| > q) \leq \min\{\Pr(|A| > q), \Pr(|B| > q)\},$$

it suffices to show that $\Pr(A > q)$ and $\Pr(B > q)$ cannot both converge to one along some subsequence of n.

Suppose, on the contrary, that $\Pr(A > q)$ and $\Pr(B > q)$ do converge to one along some subsequence of n (retain notation). The proof proceeds in two steps. In the first step we show that for large n both λ_A and λ_B are strictly above ν. Moreover, if either λ_A or λ_B converges to ν, then it converges at a rate slower than $\frac{1}{\sqrt{n}}$. In the second step we show that this implies that the equilibrium cutoffs, u_A and u_B, must be growing to infinity, in contradiction to step 1.

Step 1. $\lim_{n\to\infty} \frac{(\lambda_A - \nu)\sqrt{n}}{\sqrt{\lambda_A(1-\lambda_A)}} = \infty$ and $\lim_{n\to\infty} \frac{(\lambda_B - \nu)\sqrt{n}}{\sqrt{\lambda_B(1-\lambda_B)}} = \infty$.

We prove $\lim_{n\to\infty} \frac{|\lambda_A - \nu|\sqrt{n}}{\sqrt{\lambda_A(1-\lambda_A)}} = \infty$; similar arguments hold for λ_B.

Assume to the contrary that there exists a subsequence for which $\lim_{n\to\infty} \frac{(\lambda_A^{k_n} - \nu)\sqrt{n}}{\sqrt{\lambda_A(1-\lambda_A)}} = c$, where $-\infty \leq c < \infty$.

Let X_n denote the number of A announcements (i.e., $|A|$). By the Berry-Esséen Theorem (see, for example, Feller, 1986, Chapter XVI.5, Theorem 1), for some $\varepsilon < \Phi(-c)$, there exists an N such that for $n > N$

$$\Pr(X_n > q) = \Pr\left(\frac{X_n - n\lambda_A^{k_n}}{\sqrt{n\lambda_A^{k_n}(1-\lambda_A^{k_n})}} > \frac{-(\lambda_A^{k_n} - v)\sqrt{n}}{\sqrt{\lambda_A^{k_n}(1-\lambda_A^{k_n})}} \right) < 1 - \Phi(-c) + \varepsilon < 1$$

and this contradicts our premise that $\lim_{n\to\infty}\Pr(|A| > q) = 1$.

Recalling that $\pi = \lambda_B/(1 - \lambda_A)$ and $\sigma = \lambda_A/(1 - \lambda_B)$, it follows from step 1 that $\lim_{n\to\infty} \frac{(\pi - \frac{v}{1-v})\sqrt{n}}{\sqrt{\pi(1-\pi)}} = \infty$ and $\lim_{n\to\infty} \frac{(\sigma - \frac{v}{1-v})\sqrt{n}}{\sqrt{\sigma(1-\sigma)}} = \infty$.

Step 2. If $\lim_{n\to\infty} \frac{(\pi - \frac{v}{1-v})\sqrt{m-1}}{\sqrt{\pi(1-\pi)}} = \infty$ and $\lim_{n\to\infty} \frac{(\sigma - \frac{v}{1-v})\sqrt{m-1}}{\sqrt{\sigma(1-\sigma)}} = \infty$, then $u_A \to \infty$ and $u_B \to \infty$.

As in step 1 we provide a proof for u_A and similar arguments follow for u_B.

Let Y_n be the sum of successes from a binomial distribution with probability of success π and with $m - 1$ draws. Then

$$\sum_{k=0}^{q} \binom{m-1}{k} \pi^k (1-\pi)^{m-1-k} = \Pr(Y_n \le q) \le \Pr(|Y_n - (\mathrm{m}-1)\pi| \ge (\mathrm{m}-1)\pi - q)$$

$$< \frac{Var(Y_n)}{((m-1)\pi - q)^2} = \frac{1}{\left(\frac{(\pi - \frac{q}{m-1}\sqrt{m-1}}{\sqrt{\pi(1-\pi)}} \right)^2} \to 0,$$

where the last inequality is by Chebyshev's inequality and the limit follows from the premise. Therefore, by (5) it must be that $u_A \to \infty$. This implies that $\lambda_A \to 0$, in contradiction to step 1. ■

Proof of Observation 5. In place of the variable u, define a variable u^a for the A types by

$$u^a \equiv \alpha \frac{v - v'}{v'},$$

and a corresponding variable u^b for the B types by

$$u^b \equiv (1-\alpha)\frac{v - v'}{v'}.$$

Nothing changes in our description of the equilibrium conditions (5) and (6), except that a Z-type defines her threshold u_Z using the variable u_z. Notice that the cdfs of u^a and u^b are now different, but that

$$F^a(u^a) = F\left(\frac{u^a}{2\alpha}\right) \text{ and } F^b(u^b) = F\left(\frac{u^b}{2[1-\alpha]}\right). \tag{44}$$

Now, suppose that along some sequence α converging to 1, λ_B does *not* converge to zero. Then, because $\lambda_B = (1-p)[1-F^b(u_B)]$, $F^b(u_B)$ fails to converge to 1, which means (using (44)) that u_B must converge to zero. Using (6), we must conclude that

$$\sigma = \frac{\lambda_A}{1-\lambda_B}$$

converges to 0. But this must imply in turn that λ_A converges to 0, or that $F^a(u_A)$ converges to 1. Using (44) again, we must conclude that $u_A \to \infty$, so by (5),

$$\pi = \frac{\lambda_B}{1-\lambda_A}$$

converges to 1. With λ_A converging to 0 and λ_B bounded above by $1-p = 1/2$, this is an impossibility.

So we have shown that λ_B converges to zero. Because λ_A is bounded above by $p = 1/2$, this means that

$$\pi = \frac{\lambda_B}{1-\lambda_A}$$

converges to zero as well. An inspection of (5) now shows that u_A must converge to 0. Using (44), it follows that $F^a(u_A)$ also converges to 0, which proves that $\lambda_A = p[1-F^a(u_A)]$ converges to $p = 1/2$. ■

Proof of Observation 6. To prove part [1], define $\delta \equiv 1/(n-1-2q)$, and rewrite (12) as

$$(1 + w_B^\delta)[1 - H(w_B)] = 1/(1-p). \tag{45}$$

Notice that when $w_B = 1$, the LHS of (45) equals 2, while the RHS is strictly smaller than 2 (because $p = 1/2$).

Now suppose that there is some w such that the LHS of (45), evaluated at $w_B = w$, is strictly less than $1/(1-p)$. In this case, consider some intersection $x = w_B$ of the function $(1 + x^\delta)[1 - H(x)]$ with the value $1/(1-p)$, along with the value $w_A = 1$. It can be verified that such an intersection constitutes a semi-corner minority equilibrium.

It remains to show that the condition in the first line in the previous paragraph is satisfied for all (n, q) large enough. To this end, fix some w such that $1 - H(w) < ½(1-p)$. Now take (n, q) to infinity and notice that $\delta \to 0$. Therefore w^δ converges to 1. It follows that for large (n, q),

$$(1 + w^{\delta})[1 - H(w)] < 1/(1 - p),$$

and we are done.

Note that part [2] is trivially true for corner minority equilibria. To prove part [2] for semi-corners, note that the probability that the minority outcome is implemented is given by

$$\Pr(|A|) \geq m) = \sum_{k=m}^{n} \binom{n}{m} [p + (1-p)H(w_B)]^k [(1-p)(1-H(w_B))]^{n-k}$$

Similarly,

$$\Pr(|B|) \geq m) = \sum_{k=m}^{n} \binom{n}{m} [(1-p)(1-H(w_B))]^k [p + (1-p)H(w_B)]^{n-k}$$

Thus, $\Pr(|A| \geq m) > \Pr(|B| \geq m)$ if and only if $(1 - p)(1 - H(w_B)) < p + (1 - p)H(w_B)$, which may be rewritten as

$$\frac{1}{2(1-p)} > 1 - H(w_B) \tag{46}$$

Now (45) tells us that

$$1 - H(w_B) = \frac{1}{(1-p)(1+w_B^{\delta})}$$

where $w_B > 1$. Hence, $(1 - p)(1 + \mathrm{w}_B^{\delta}) > 2(1 - p)$, which implies (46). ■

Proof of Observation 7. Let w_B^* be the solution to the following equation:

$$p + (1 - p)H(w_B^*) = (1 - p)[1 - H(w_B^*)]$$

Notice that w_B^* is well-defined and greater than 1, as long as p = ½. We now proceed in two steps.

Step 1. There exists a sequence of semi-corner minority equilibria that converges to $(1, w_B^*)$. To see this, note that when $w_B = w_B{}^*$ the RHS of (13) is smaller than the LHS. For any $\varepsilon > 0$, set $w_B = w_B{}^* + \varepsilon$. Because $\frac{p+(1-p)H(w_B+\varepsilon)}{(1-p)[1-H(w_B{}^*+\varepsilon)} > 1$ there exists $N(\varepsilon) < \infty$ such that for all $n \geq N(\varepsilon)$, the LHS of (13) is strictly greater than its RHS. It follows that for all $n \geq N(\varepsilon)$, there exists an equilibrium $(1, w_B^n)$ where $w_B^n \in (w_B^*, w_B^* + \varepsilon)$.

Step 2. By Step 1, as $n \to \infty$, the probabilities with which a random voter votes for A or for B (along the above sequence of semi-corner minority equilibria) both converge to 1/2. In particular, there exists an N above which these probabilities are bounded below by $\bar{\nu} > \nu$ and above by $1 - \bar{\nu}$. The probability of disagreement is equal to $1 - \Pr(|A| \geq m) - \Pr(|B| \geq m)$. We now show that $\Pr(|A| \geq m)$ goes to zero as $n \to \infty$. By essentially the same argument, $\Pr(|B| \geq i)$ also goes to zero as $n \to \infty$.

Recall that

$$\Pr(|A|) \geq m) = \sum_{k=m}^{n} \binom{n}{m} [p + (1-p)H(w_B)]^k [(1-p)(1-H(w_B))]^{n-k}$$

Note that $\left|\frac{m}{n} - (1-\nu)\right| < \frac{1}{n}$. Because $1 - \bar{\nu} < 1 - \nu$ it follows that for large enough n,

$$1 - \bar{\nu} < \frac{m}{n} - \eta \tag{47}$$

for some $\eta > 0$. By stochastic dominance,

$$\Pr(|A|) \geq m) \leq \sum_{k=m}^{n} \binom{n}{m} (1-\bar{\nu})^k (\bar{\nu})^{n-k} \tag{48}$$

By inequality (47) and the SLLN, the RHS of (48) goes to zero. ■

Observation 8. *Consider the model with a majority tie-breaking rule. All symmetric equilibria in this model are interior.*

Proof. We proceed in three steps.

Step 1. No side can use an infinite cutoff in equilibrium. Suppose that side A does. Then note that no matter what rule side B follows, $\tau + \tau' > 0$. This is because the sum of probabilities $\tau + \tau'$ is greater than the probability that both sides have exactly zero votes, which in turn, is at least as high as the probability that all individuals but one are A types. Since the latter probability is positive we obtain the desired inequality.

In order for an A-type to declare neutrality in equilibrium, his u value must satisfy $(\tau + \tau')u < P^+$. But this inequality cannot hold for an infinite u because we have just shown that $\tau + \tau' > 0$, a contradiction. Hence, equilibrium cutoffs of both sides are bounded.

Step 2. $P^+ > 0$. To prove this, fix some person, say of type A, and simply take the event in which exactly q compatriots are of type A (apart from the special individual) and the rest are of type B, and all value realizations are above the cutoffs. Because cutoffs are bounded, the probability of this event is strictly positive. But this event is contained within the one covered by P^+. So $P^+ > 0$.

Step 3. In any equilibrium, cutoffs are strictly positive. To see this, note by step 2 that $P^+ > 0$. Now take u very small; the inequality $(\tau + \tau')u \geq P^+$ cannot hold. ■

References[21]

Araki, K. and T. Börgers (1996), 'How minorities can win majority votes', mimeo, Department of Economics, University College London.

Austen-Smith, D. and T. Feddersen (2002), 'The inferiority of deliberation under unanimity', mimeo, Kellogg Graduate School of Management, Northwestern University.

Buchanan, J. and G. Tullock (1962), *The Calculus of Consent*, Ann Arbor: University of Michigan Press.

Campbell, C. (1999), 'Large electorates and decisive minorities', *Journal of Political Economy* **107**, 1199–1217.

Casella, A. (2005), 'Storable votes', *Games and Economic Behavior* **51**, 391–419.

Cason, T. and V. Mui (1997), 'A laboratory study of group polarization in the team dictator game', *Economic Journal* **107**, 1465–1483.

Coughlan P. (2000), 'In defense of unanimous jury verdicts: Mistrials, communication, and strategic voting', *American Political Science Review* **94**(2), 375–393

Dodgson, C. (1984), *The Principles of Parliamentary Representation*, supplement, 1885, London: Harrison and Sons.

Feller, W. (1986), *An Introduction to Probability Theory and its Applications*, Volume II, New Delhi: Wiley Eastern Limited.

Gerardi, D. and L. Yariv (2003), 'Putting your ballot where your mouth is – An analysis of collective choice with communication', mimeo, Department of Economics, Yale University.

Gerber, E., R. Morton and T. Rietz (1998), 'Minority representation in multi-member districts', *American Political Science Review* **92**, 127–144.

Goeree, J. and J. Grosser (2005), 'False consensus voting and welfare reducing polls', mimeo, Department of Economics, University of Cologne.

Haan, M. and P. Kooreman (2003), 'How majorities can lose the election: Another voting paradox', *Social Choice and Welfare* **20**, 509–522.

Hortala-Vallve, R. (2004), 'Qualitative voting', mimeo, Department of Economics, London School of Economics.

Jackson, M. and H. Sonnenschein (2007), 'Overcoming incentive constraints by linking decisions', *Econometrica* **75**(1), 241–257.

Krasa, S. and M. Polborn (2004), 'Is mandatory voting better than voluntary voting?', mimeo, Kellogg Graduate School of Management, Northwestern University.

Lamm, H. and D. Myers (1978), 'Group-induced polarization of attitudes and behavior', in L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, vol. 11, New York: Academic Press, pp. 145–195.

Ledyard, J. (1984), 'The pure theory of large two candidate elections', *Public Choice* **44**, 7–41.

Lijphart, A. (1999), *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*, New Haven and London: Yale University Press.

Myers, D. and H. Lamm (1976), 'The group polarization phenomenon', *Psychological Bulletin* **83**, 602–627.

The New York Times (2003), 'Younger sniper gets a sentence of life in prison', Dec 24.

Olson, M. (1965), *The Logic of Collective Action: Public Goods and the Theory of Groups*, Cambridge, MA: Harvard University Press.

[21] [Editor's note] Jackson and Sonnenschein (2007), which was originally cited as forthcoming, has been updated.

Palfrey, T. and H. Rosenthal (1983), 'A strategic calculus of voting', *Public Choice* **41**, 7–53.

Pareto, V. (1906 [1927]), *Manual of Political Economy*, New York: A.M. Kelley.

Philipson, T. and J. Snyder (1996), 'Equilibrium and efficiency in an organized vote market', *Public Choice* **89**, 245–265.

Piketty, T. (1999), 'Information Aggregation through Voting and Vote Trading', mimeo, Department of Economics, Massachusetts Institute of Technology.

Ponsati, C. and J. Sàkovics (1996), 'Multiperson bargaining over two alternatives', *Games and Economic Behavior* **12**, 226–244.

Sunstein, C. (2000), 'Deliberative trouble? Why groups go to extremes', *Yale Law Journal* **110**, 71–119

Sunstein, C. (2002), 'The law of group polarization', *Journal of Political Philosophy* **10**, 175–195.

Sunstein, C. (2003), *Why Societies Need Dissent*, Cambridge, MA: Harvard University Press.

Strategic Basins of Attraction, the Path Dominance Core, and Network Formation Games

Frank H. Page Jr. and Myrna Wooders

Paper presented at the CTN Workshop 2005. Reprinted from *Games and Economic Behavior* **66**(1), May 2009, 462–487, with permission from Elsevier.

We introduce a model of network formation whose primitives consist of a feasible set of networks, player preferences, rules of network formation, and a dominance relation on feasible networks. Rules may range from noncooperative, where players may only act unilaterally, to cooperative, where coalitions of players may act in concert. The dominance relation over feasible networks incorporates player preferences, the rules of network formation, and the degree of farsightedness of players. A specification of the primitives induces an abstract game consisting of (i) a feasible set of networks, and (ii) a path dominance relation. Using this induced game we characterize sets of network outcomes that are likely to emerge and persist. Finally, we apply our approach and results to characterize the equilibrium of well known models and their rules of network formation, such as those of Jackson and Wolinsky (1996) and Jackson and van den Nouweland (2005).

1. Introduction

1.1 Overview of the questions, the model and the main results

In many economic and social situations the totality of interactions between

We are especially grateful to Michael Maschler and Salvador Barbera for stimulating discussions and for an example. We also wish to thank John Conlon, Anne van den Nouweland and participants at various conferences where this paper was presented for their interest and comments. Finally, we wish to thank two anonymous referees whose careful reading and thoughtful comments led to many improvements in the paper. This paper continues our earlier work presented at the CTN Workshop on Coalitions and Networks held at the University of Warwick in June of 2001. The current version of the paper was substantially completed while both authors were visiting CERMSEM at Paris 1 during the summers of 2005 and 2006. Both authors are grateful to CERMSEM and Paris 1 for support and hospitality.

individuals and coalitions can be modeled as a network. We address the following question: given preferences of individuals and rules governing network formation, what networks are likely to emerge and persist? To address this question we introduce a model of network formation whose primitives consist of a feasible set of networks, player preferences, the rules of network formation, and a dominance relation. The rules of network formation may range from noncooperative, where players may only act unilaterally, to fully cooperative, where coalitions consisting of multiple players may act in concert. The dominance relation may be either direct or indirect. Under direct dominance players are concerned with immediate consequences of their network formation strategies whereas under indirect dominance players are farsighted and consider the eventual consequences of their strategies. As we will discuss, our framework can accommodate a wide variety of social and economic situations.[1]

A specification of the primitives induces an abstract game consisting of (i) a feasible set of networks and (ii) a path dominance relation defined on the feasible set of networks.[2] Under the path dominance relation, a network G path dominates another network G' if there is a finite sequence of networks, beginning with G and ending with G' where each network along the sequence dominates its predecessor.[3] Using this induced abstract game as our basic analytic tool we demonstrate that for any set of primitives the following results hold:

1. The feasible set of networks contains a unique, finite, disjoint collection of nonempty subsets each constituting a *strategic basin of attraction*. Given preferences and the rules of governing network formation, these basins of attraction are the absorbing sets of the process of network formation modeled via the game.
2. A stable set (in the sense of von Neumann Morgenstern) with respect to path dominance consists of one network from each basin of attraction.

[1] Our framework is essentially that of Chwe (1994) but applied to networks. Using Chwe's framework we are able to take into account both rules and preferences in the formation of networks.

[2] To our knowledge, there are no prior papers formulating the problem of network formation as an abstract game. Because our abstract game is induced from the Chwe primitives (preferences and effectiveness relations expressing the rules of network formation) our approach is very much in the spirit of Chwe (1994) and other papers such as Gillies (1959), Harsanyi (1974), Inarra, Kuipers, and Olaizola (2005), Kalai and Schmeidler (1977), Moulin and Peleg (1982), Rosenthal (1972), and Shenoy (1980).

[3] Stated formally, given feasible set of networks $\mathbb{G}$ and dominance relation $>$, network $G' \in \mathbb{G}$ (weakly) path dominates network $G \in \mathbb{G}$, written $G' \geq_p G$, if $G' = G$ or if there exists a *finite* sequence of networks $\{G_k\}_{k=0}^{n}$ in $\mathbb{G}$ with $G = G_0$ and $G' = G_n$ such that for $k = 1, 2, \ldots, n$

$$G_k > G_{k-1}.$$

The path dominance relation $\geq_p$ induced by the dominance relation $>$ is sometimes referred to as the transitive closure of $>$.

3. The path dominance core, defined as a set of networks having the property that no network in the set is path dominated by any other feasible network, consists of one network from each basin of attraction containing a *single* network. Note that the path dominance core is contained in each stable set and is nonempty if and only if there is a basin of attraction containing a single network.[4] As a corollary, we conclude that any network contained in the path dominance core is constrained Pareto efficient. Thus, by considering the network formation game with respect to path dominance – and thus by considering the long run – we identify networks that, given the rules of network formation, are *both* stable and Pareto-efficient with respect to the original dominance relation.
4. From the results above it follows that if the dominance relation is transitive and irreflexive, then the path dominance core is nonempty.

We also demonstrate specializations of our model to treat network formation games over linking networks as well as hedonic games and we discuss how our results apply to these examples.

There are interesting connections between our notions of stability (basins of attraction, path dominance stable sets, and path dominance core) and some of the basic notions of stability and equilibrium found in the existing literature – such as, strong stability (Jackson and van den Nouweland, 2005), pairwise stability (Jackson and Wolinsky, 1996), consistency (Chwe, 1994), and Nash equilibrium. We show that in general (for all primitives) the path dominance core is contained in the set of strongly stable networks. We conclude from our general results therefore that, for all primitives, the existence of at least one basin of attraction containing a single network is sufficient for the existence of a strongly stable network. We also demonstrate that, depending on how we specialize the primitives of the model, the path dominance core is equal to the set of strongly stable networks, the set of pairwise stable networks, or the set of Nash networks.

Of particular interest are the connections between the rules of network formation, the dominance relation inducing path dominance, and stability.[5] We provide a unified and systematic analysis of these connections. For example, we show that:

(a) If path dominance is induced by a direct dominance relation (as opposed to an indirect dominance relation as in Chwe, 1994, for example), then the path dominance core is equal to the set of strongly stable networks.

[4] Put differently, the path dominance core is empty if and only if all basins of attraction contain multiple networks.

[5] Although she treats a more specialized model, the questions addressed in Demange (2004) are related.

(b) If, in addition, the rules of network formation are the Jackson-Wolinsky rules, then the path dominance core is equal to the set of pairwise stable networks.[6]

(c) If path dominance is induced by a direct dominance relation and if the rules of network formation only allow network changes brought about by individuals, then the path dominance core is equal to the set of Nash networks.

We then conclude from (3) above, the existence of at least one basin of attraction containing a single network is, depending on how we specialize primitives, both necessary and sufficient for either (i) the existence of a strongly stable network, or (ii) a pairwise stable network, or (iii) a Nash network.[7]

For path dominance induced by an indirect dominance relation as in Chwe (1994), we show that for all primitives – and in particular for all rules of network formation – each strategic basin of attraction has a nonempty intersection with the largest consistent set of networks (i.e., the Chwe set of networks, see Chwe, 1994).[8] This fact, together with (2) above, implies that there always exists a path dominance stable set contained in the largest consistent set. Thus, the path dominance core is contained in the largest consistent set. In light of our results on the path dominance core and stability (both strong and pairwise), we conclude that if path dominance is induced by an indirect dominance relation, then any network contained in the path dominance core is not only consistent but also strongly stable, as well as pairwise stable.[9]

We remark that solution concepts defined using dominance relations have a distinguished history in the literature of game theory. First, consider the von-Neuman-Morgenstern stable set. The vN-M stable set is defined with respect to a dominance relation on a set of outcomes and consists of those outcomes that are externally and internally stable with respect to the given dominance relation.[10] Similarly, Gilles (1959) defines the core based on a given dominance relation. These solution concepts, with a few exceptions, have typically been applied to

[6] Under the Jackson-Wolinsky rules arc addition is bilateral (i.e., the two players that would be involved in the arc must agree to adding the arc), arc subtraction is unilateral (i.e., at least one player involved in the arc must agree to subtract or delete the arc), and network changes take place one arc at a time (i.e., in any one play of the game, only one arc can be added or subtracted). See section 3.2.1 for a formal definition.

[7] For Jackson-Wolinsky linking networks, Calvó-Armengol and Ilkilic (2005) provide necessary and sufficient conditions on the network link marginal payoffs such that the set of pairwise stable, pairwise Nash, and proper equilibrium networks coincide.

[8] Consistency with respect to indirect dominance and the notion of a largest consistent set were introduced by Chwe (1994) in an abstract game setting. We provide a detailed discussion of Chwe's notion in Section 5.3.

[9] Other papers on indirect dominance and consistency in games include Xue (1998), Diamantoudi and Xue (2003), and Mauleon and Vannetelbosch (2003).

[10] Richardson (1953) gives properties an irreflexive dominance relation must satisfy relative to a given set of outcomes in order to guarantee the existence of a vN-M stable set.

models of economies or cooperative games where the notion of dominance is based on what a coalition can achieve using only the resources owned by its members (cf., Aumann, 1964) or a given set of utility vectors for each possible coalition (cf., Scarf, 1967). Particularly notable exceptions are Schwartz (1974), Kalai et al. (1976), Kalai and Schmeidler (1977) and Shenoy (1980). Their motivations are in part similar to ours in that they take as given a set of possible choices of a society and a dominance relation and, based on these, describe a set of possible or likely social outcomes called, by Kalai and Schmeidler, the admissible set. While their examples treat direct dominance, their general results have wider applications. We return to a discussion of the admissible set in our concluding section.

1.2 A further discussion of the model

In addition to introducing abstract games of network formation, our modeling approach contributes to the literature by extending the class of primitives used in the analysis of network formation in three respects. These extensions, listed below, significantly broaden the set of potential applications.

1.2.1. Directed networks with heterogenous arcs and multiple uses of arcs of the same type

First, we focus on directed networks rather than on linking networks[11] and distinguish between nodes and decision making players (i.e., the set of players and the set of nodes are not necessarily the same). Connections are represented by arcs and each arc possesses an orientation or direction: arc a connecting nodes i and i' must either go from node i to node i' or must go from node i' to node i.[12] For example, an individual may have links on his web page to the web pages of all Nobel Laureates in economics but it may be that no Nobel Laureate has a link to that individual's web page. Connections between nodes (i.e., arcs), besides having an orientation, are allowed to be heterogeneous. To illustrate, if the nodes in a given network represent players, an arc a going from player i to player i' might represent a particular type and intensity of interaction (identified by the arc label a) initiated by player i towards player i'. Player i might direct great affection toward player i' as represented by arc type a, but player i' may direct only lukewarm affection toward player i as represented by arc type a'.

Under our extended definition nodes are allowed to be connected by multiple, distinct arcs. Thus, we allow nodes to interact in multiple, distinct ways. For example, nodes i and i' might be connected by arcs a and a', with arc a running from node i to i' and arc a' running in the opposite direction (i.e., from

[11] In particular, we focus on the notion of directed networks introduced in Page et al. (2005).

[12] We denote arc a going from node i to node i' via the ordered pair $(a, (i, i'))$, where (i, i') is also an ordered pair. Alternatively, if arc a goes from node i' to node i, we write $(a, (i', i))$.

node i' to node i).[13] If node i represents a seller and node i' a buyer, then arc a might represent a contract offer by the seller to the buyer, while arc a' might represent a counter offer or the acceptance or rejection of the contract offer. Finally, loops are allowed and arcs are allowed to be used multiple times in a given network.[14] For example, arc a might be used to connect nodes i and i' as well as nodes i' and i''. Thus, under our definition nodes i and i' as well as nodes i' and i'' are allowed to engage in the same type of interaction as represented by arc type a.

Allowing each type of arc to be used multiple times makes it possible to distinguish coalitions by the type of interaction taking place between coalition members and to give a network representation of such coalitions. For example, if the nodes in a given network represent players, an 'a-coalition' could consist of all players i having an a-connection with at least one other player i'. Such an a-coalition would then have a network representation as the directed subnetwork consisting of pairs of nodes, i and i', connected by arc type a.

Until now, most of the economic literature on networks has focused on linking networks (see Jackson, 2005 for an excellent survey). In an undirected (or linking) network, an arc (or link) is identified with a nonempty subset of nodes consisting of exactly two distinct nodes, for example, $\{i, i'\}$, $i \neq i'$. Thus, in an undirected network, a link has no orientation and simply indicates a connection between two players. Moreover, links are typically not distinguished by type (or by label) – that is, links are homogeneous. By allowing arcs to possess direction and the same type of arc to be used multiple times and by allowing loops and nodes to be connected by multiple arcs, our definition makes possible the application of networks to a rich set of economic environments. For example, a job opportunity market model may embody the features introduced above; individuals may have different relationships with their superiors in an organization and other individuals both within and outside of the organization. This may well affect social interactions and job opportunities.

1.2.2. The rules of network formation

We explicitly model the rules of network formation and thus provide a systematic treatment of the relationship between rules and stability. The rules of network formation specify which players must be involved in adding, subtracting, or replacing an arc as well as how many and what types of arcs can be added, subtracted, or replaced in any one play of the game.

In much of the literature, it is assumed (sometimes implicitly) that network

[13] Under our extended definition, arc a' might also run in the same direction as arc a. However, our definition does not allow arc a to go from node i to node i' multiple times.

[14] A loop is an arc going *from* a given node *to* that same node. For example, given arc a and node i, the ordered pair $(a,(i,i))$ is a loop.

formation is governed by the Jackson-Wolinsky rules.[15] Other rules are possible. For example, the addition of an arc might require that a simple majority of the players agree to the addition while the deletion of an arc might require that a two-thirds majority agree to the deletion. Under our approach, such rules are allowed. We achieve this flexibility by representing the rules of network formation via a collection of coalitional effectiveness relations, $\{\rightarrow_S\}_S$, defined on the feasible set of networks. Given feasible networks G and G', if the relation $G \rightarrow_S G'$ holds, the players in coalition S can change network G to network G'. In constructing our abstract game of network formation, we will equip the feasible set of networks with a dominance relation which incorporates – or represents – *both* the preferences of individuals *and* coalitions and the rules of network formation as represented via the coalitional effectiveness relations $\{\rightarrow_S\}_S$. Thus, the stability results we obtain using the path dominance relation will reflect both preferences and rules.

1.2.3 The dominance relation defined on feasible networks

While all of our main results (Section 4) hold for path dominance induced by any binary relation, we will focus primarily on path dominance induced by either direct dominance or indirect dominance (Sections 3.3.1 and 3.3.2).

1.3 Examples

To demonstrate the flexibility of our approach and the wide applicability of our results, we consider three examples. Our first example treats noncooperative network formation games and shows that any such network formation game possessing a potential function has basins of attraction each consisting of a single network – and thus shows that any noncooperative network formation game possessing a potential function has a nonempty path dominance core. Our second example demonstrates how our approach can be applied to Jackson-Wolinsky linking networks and provides necessary and sufficient conditions for nonemptiness of the set of pairwise stable linking networks. Finally, our third example, proposed to us by Salvador Barbera and Michael Maschler in private correspondence, shows how our framework also encompasses hedonic games – games where players' preferences are defined over the set of coalitions in which they may be members. The example illustrates how, through indirect dominance, outcomes in a game might move from one hedonic core point to another. From our prior results, this demonstrates that, even though the hedonic core, that is the

[15] Jackson-van den Nouweland (2005) focus on linking networks and assume that link addition is bilateral while link subtraction is unilateral. But in their model, network changes are not required to take place one link at a time – multiple link changes can take place in any one play of the game. We shall refer to these rules as the Jackson-van den Nouweland rules. Calvó-Armengol and Ilkilic (2004) also consider linking networks under bilateral-unilateral rules and allow multiple link changes.

core with respect to direct dominance, is nonempty, the hedonic farsighted core, that is the core with respect to indirect dominance, is empty. (In related work Diamantoudi and Xue, 2003 also investigate hedonic games with indirect dominance, but with a different set of effectiveness relations than we consider here).

2. Directed Networks

2.1 The definition

Let N be a finite set of nodes, with typical element denoted by i, and let A be a finite set of arcs types (or simply arcs), with typical element denoted by a. Arcs represent potential types of connections between nodes, and depending on the application, nodes can represent economic players or economic objects such as markets or firms. The following definition is from Page et al. (2005).

Definition 1 – Directed Networks. Given node set N and arc set A, a directed network, G, is a nonempty subset of $A \times (N \times N)$. The collection of all directed networks is denoted by $P(A \times (N \times N))$.

A directed network $G \in P(A \times (N \times N))$ specifies how the nodes in N are connected via the arcs in A. Note that in a directed network order matters. In particular, if $(a, (i, i')) \in G$, this means that arc *a goes from* node *i to* node i'. Also, note that loops are allowed – that is, we allow an arc to go from a given node back to that given node. For example, in a network model of journal citations loops could represent self-cites.[16] Finally, an arc can be used multiple times in a given network and multiple arcs can go from one node to another. However, under our definition an arc a is not allowed to go from a node i to a node i' multiple times.

The following notation is useful in describing changes in networks and the properties of networks. Given directed network $G \in P(A \times (N \times N))$, let $G \cup (a, (i, i'))$ denote the network obtained by adding arc a from node i to node i' to network G, and let $G \backslash (a, (i, i'))$ denote the network obtained by subtracting (or deleting) arc a from node i to node i' from network G. Also, let

$$\left.\begin{array}{l} G(a) := \{(i,i') \in N \times N : (a,(i,i')) \in G\}, \\ \text{and} \\ G(i) := \{a \in A : \text{for some } i' \in N \text{ either } (a,(i,i')) \in G \text{ or } (a,(i',i)) \in G\}. \end{array}\right\} \qquad (1)$$

Thus, $G(a)$ is the *set of node pairs* connected by arc a in network G, and $G(i)$ is the *set of arcs* going from node i or coming to node i in network G.

[16] Other examples could be developed. For example, in a network model of information sharing, the fact that each player knows his own information would be represented by a loop.

Note that if for some arc $a \in A$, $G(a)$ is empty, then arc a is not used in network G. Moreover, if for some node $i \in N$, $G(i)$ is empty then node i is not used in network G, and node i is said to be isolated relative to network G.

Suppose that the node set N is given by $N = \{i_1, i_2, ..., i_5\}$, while the arc set A is given by $A = \{a_1, a_2, ..., a_5, a_6, a_7\}$. Consider network G in Figure 1.

Note that in network G nodes i_1 and i_2 are connected by two a_1 arcs running in opposite directions and that nodes i_1 and i_3 are connected by two arcs, a_1 and a_3, running in the same directions from node i_3 to node i_1. Thus, $G(a_1) = \{(i_1, i_2), (i_2, i_1), (i_3, i_1)\}$ and $G(a_3) = \{(i_3, i_1)\}$. Observe that $(a_6, (i_4, i_4)) \in G$ is a loop. Thus, $G(a_6) = \{(i_4, i_4)\}$. Also, observe that arc a_7 is not used in network G. Thus, $G(a_7) = \varnothing$. Finally, observe that $G(i_4) = \{a_4, a_5, a_6\}$, while $G(i_5) = \varnothing$. Thus, node i_5 is *isolated* relative to G.[17]

Throughout we shall take as the feasible set of networks some nonempty subset $\mathbb{G}$ of $P(A \times (N \times N))$.

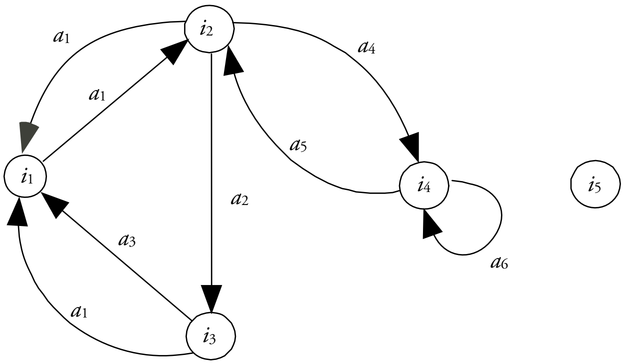


Figure 1: Network G

2.2 Linking networks, directed graphs, and directed networks

As before, let N denote a finite set of nodes. A linking network, say g, consists of a finite collection of subsets of the form $\{i, i'\}$, $i \neq i'$. Thus, $\{i, i'\} \in g$ means that nodes i and i' are linked in network g. For example, g might be given by $g = \{\{i, i'\}, \{i', i''\}\}$ for i, i', and i'' in N. Note that all connections or links are the same (i.e., connection types are homogeneous), direction does not matter, and loops are ruled out. Letting g^N denote the collection of all subsets of N of size 2, the collection of all linking networks given N is given by $P(g^N)$ where $P(g^N)$ denotes the collection of all nonempty subsets of g^N (e.g., see the definition in Jackson and Wolinsky, 1996).[18]

[17] If the loop $(a_7, (i_5, i_5))$ were part of network G in Figure 1, then node i_5 would no longer be considered isolated under our definition. Moreover, we would have $G(i_5) = \{a_7\}$.

[18] In section 6.3, we show how our approach to network formation games, as well as some of our main results, can be applied to linking networks.

A directed graph, say E, consists of a finite collection of ordered pairs $(i, i') \in N \times N$. For example, E might be given by $E = \{(i, i'), (i', i')\}$ for (i, i') and (i', i') in $N \times N$. Stated more compactly, a directed graph E is simply a subset of $N \times N$. Thus, in any directed graph connection types are again homogeneous but direction does matter and loops are allowed.

Under our definition, a directed network G is a subset of $A \times (N \times N)$, where as before A is a finite set of arcs. Thus, in a directed network, say $G \in P(A \times (N \times N))$, connection types are allowed to be heterogeneous (distinguished by arc labels), direction matters, and loops are allowed.

Formally, linking networks are not a special cases of directed networks. However, any linking network can be given an alternative representation as a directed network. To see this, consider linking network $g \in P(g^N)$ and suppose nodes i and i' are linked in network g (i.e., $\{i, i'\} \in g$). Next consider a directed network $G \in P(A \times (N \times N))$ where the set of arc types A contains one arc, $A = \{1\}$, and say that nodes i and i' are directly linked in G if and only if there is an arc from i to i' and another arc from i' to i.[19] We say that directed network G is an alternative representation of linking network g provided

$\{i, i'\} \in g$ if and only if i and i' are directly linked in G.

With multiple arc types, directed networks allow us to differentiate links by types or intensity levels, and thus allow us to consider a richer collection of links between nodes. For example, suppose that A contains multiple arc types each specifying a type of connection or an intensity level of a connection. We say that i and i' are a-linked in network $G \in P(A \times (N \times N))$ provided both $(a, (i, i'))$ *and* $(a, (i', i))$ are in G. Thus, various sorts of links between players can be modelled and analyzed.

As we shall show in Section 6.2, in addition to the fact that linking networks can be given alternative representations as directed networks, the game theoretic approach to network formation we shall develop here can be applied directly to linking networks.

3. Preferences, Rules, and Dominance Relations

3.1 Preferences

Let D denote the set of players (or economic decision making units) with typical element denoted by d, and let $P(D)$ denote the collection of all coalitions (i.e., nonempty subsets of D) with typical element denoted by S. Note that, the set of players D and the set of nodes N are not necessarily the same set.

[19] Thus, nodes i and i' are directly linked in G if and only if $(1, (i, i'))$ and $(1, (i', i))$ are in G. Whereas, nodes i and i' are connected if and only if $(1, (i, i'))$ or $(1, (i', i))$ is in G (i.e., mutual arcs raise a connection to the level of a link).

Given a feasible set of directed networks $\mathbb{G} \subseteq P(A \times (N \times N))$, we shall assume that each player's preferences over networks in $\mathbb{G}$ are specified via an *irreflexive* binary relation $\succ_d$. Thus, player $d \in D$ prefers network $G' \in \mathbb{G}$ to network $G \in \mathbb{G}$ if $G' \succ_d G$ and for all networks $G \in \mathbb{G}$, $G \nsucc_d G$ (irreflexivity). Coalition $S' \in P(D)$ *prefers network G' to network G*, written $G' \succ_{S'} G$, if $G' \succ_d G$ for all players $d \in S'$.

In many applications, a player's preferences are specified via a real-valued network payoff function, $v_d(\cdot)$. For each player $d \in D$ and each directed network $G \in \mathbb{G}$, $v_d(G)$ is the payoff to player d in network G. Player d then prefers network G' to network G if $v_d(G') > v_d(G)$. Moreover, coalition $S' \in P(D)$ prefers network G' to network G if $v_d(G') > v_d(G)$ for all $d \in S'$. Note that the payoff $v_d(G)$ to player d depends on the entire network. Thus, the player may be affected by directed links between other players even when he himself has no direct or indirect connection with those players. Intuitively, 'widespread' network externalities are allowed.

Remark 1. All of our results on basins of attraction, path dominance stable sets, and the path dominance core (Theorems 1–4 below) remain valid even if coalitional preferences $\{\succ_S\}_{S \in P(D)}$ over networks are based on weak preference relations $\{\succsim_d\}_{d \in D}$. If $G' \succsim_d G$ then player d either strictly prefers G' to G (denoted $G' \succ_d G$) or is indifferent between G' and G (denoted $G' \sim_d G$). Given weak preferences $\{\succsim_d\}_{d \in D}$, coalition $S' \in P(D)$ *prefers* network G' to network G, written $G' \succ_{S'} G$, if for all players $d \in S'$, $G' \succsim_d$ G and if *for at least one player $d' \in S'$*, $G' \succ_{d'} G$. Note that if coalitional preferences $\{\succ_S\}_{S \in P(D)}$ are defined in this way (i.e., using weak preferences $\{\succsim_d\}_{d \in D}$, then they are irreflexive (i.e., $G \nsucc_S G$ for all $G \in \mathbb{G}$ and $S \in P(D)$).

3.2 Rules

The rules of network formation are specified via a collection of coalitional effectiveness relations $\{\rightarrow_S\}_{S \in P(D)}$ defined on the feasible set of networks $\mathbb{G}$. Each effectiveness relation $\rightarrow_S$ represents what a coalition S can do. Thus, if $G \rightarrow_S G'$ this means that under the rules of network formation coalition $S \in P(D)$ can change network $G \in \mathbb{G}$ to network $G' \in \mathbb{G}$ by adding, subtracting, or replacing arcs in G.

3.2.1 Examples of Network Formation Rules

Jackson-Wolinsky Rules: To illustrate, consider Figure 2 depicting two networks G_1 and G_2 in which the nodes represent players. Thus, $D = N = \{i_1, i_2, i_3\}$.

Observe that

$$G_2 = G_1 \cup (a_1, (i_3, i_1)) \quad \text{and} \quad G_1 = G_2 \backslash (a_1, (i_3, i_1)).$$

Assume that

(i) adding an arc a from player i to player i' requires that both players i *and* i' agree to add arc a (i.e., arc addition is bilateral);

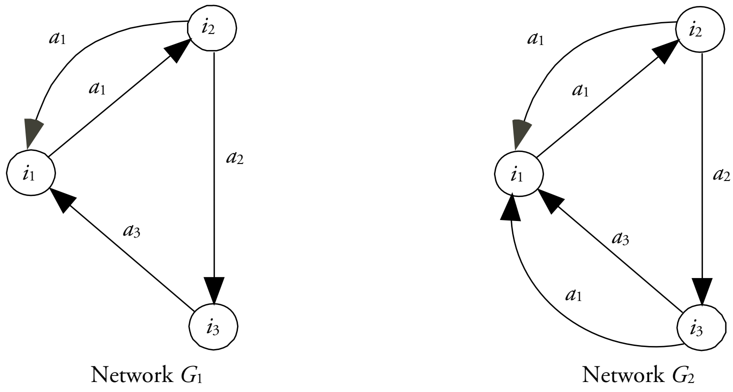


Figure 2

(ii) subtracting an arc a from player i to player i' requires that player i or player i' agree to subtract arc a (i.e., arc subtraction is unilateral);

(iii) for any pair of networks G and G' in $\mathbb{G}$, if $G \rightarrow_S G'$, then $G \neq G'$ and

either $G' = G \cup (a, (i, i'))$ for some $(a, (i, i')) \in A \times (N \times N)$
or
$G' = G \backslash (a, (i, i'))$ for some $(a, (i, i')) \in A \times (N \times N)$.

For the case $D = N$ (i.e., players = nodes), we shall refer to rules (i)–(iii) above as Jackson-Wolinsky rules. Note that rules (i) and (ii) imply that if $G \rightarrow_S G'$, then $1 \leq |S| \leq 2$. Referring to Figure 2, the effectiveness relations over networks G_1 and G_2 under Jackson-Wolinsky rules are given by

$$G_1 \xrightarrow[\{i_1, i_3\}]{} G_2 \quad G_2 \xrightarrow[\{i_1, i_3\}]{} G_1 \quad G_2 \xrightarrow[\{i_1\}]{} G_1 \quad G_2 \xrightarrow[\{i_3\}]{} G_1.$$

Jackson-van den Nouweland rules: Consider networks G_0 and G_3 depicted in Figure 3 and again suppose that nodes represent players.

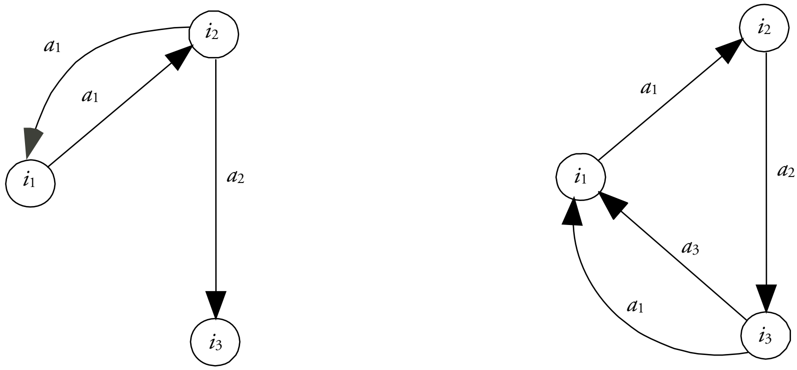
Observe that

$G_3 = (G_0 \backslash (a_1, (i_2, i_1))) \cup (a_1, (i_3, i_1)) \cup (a_3, (i_3, i_1))$
and
$G_0 = (G_3 \backslash ((a_1, (i_3, i_1)) \cup (a_3, (i_3, i_1)))) \cup (a_1, (i_2, i_1))$.

Assume that

(i) adding an arc a from player i to player i' requires that both players i *and* i' agree to add arc a (i.e., arc addition is bilateral);

(ii) subtracting an arc a from player i to player i' requires that player i *or* player i' agree to subtract arc a (i.e., arc subtraction is unilateral);



Network G_0 Network G_3

Figure 3

For the case $D = N$ (i.e., players = nodes), we shall refer to rules (i)-(ii) above as Jackson-van den Nouweland rules. Thus, the Jackson-van den Nouweland rules are the Jackson-Wolinsky rules without the one-arc-at-a-time restriction. Note that if arc addition is bilateral and arc subtraction is unilateral (i.e., if rules (i) and (ii) hold), then $G \to_S G'$ implies that G' is obtainable from G via coalition S, that is,

(i) $(a, (i, i')) \in G'$ and $(a, (i, i')) \notin G$
$\Rightarrow \{i, i'\} \subseteq S$;

(ii) $(a, (i, i')) \notin G'$ and $(a, (i, i')) \in G$
$\Rightarrow \{i, i'\} \cap S \neq \varnothing$.

Referring to Figure 3, the effectiveness relations over networks G_0 and G_3 under Jackson-van den Nouweland rules are given by

$$G_0 \xrightarrow[\{i_1,i_2,i_3\}]{} G_3 \quad G_0 \xrightarrow[\{i_1,i_3\}]{} G_3 \quad G_3 \xrightarrow[\{i_1,i_2\}]{} G_0 \quad G_2 \xrightarrow[\{i_1,i_2,i_3\}]{} G_1.$$

Noncooperative Rules: Again suppose that nodes represent players and assume that

(i) adding an arc a from player i to player i' requires only that player i agree to add the arc (i.e., arc addition is unilateral and can be carried out only by the initiator, player i);

(ii) subtracting an arc a from player i to player i' requires only that player i agree to subtract the arc (i.e., arc subtraction is unilateral and can be carried out only by the initiator, player i);

(iii) $G \to_S G'$ implies that $|S| = 1$ (i.e., only network changes brought about by individual players are allowed).

We shall refer to rules (i)–(iii) as noncooperative. Note that a player i can add or subtract an arc to player i' without regard to the preferences of player i'. Thus in general under noncooperative rules, effectiveness relations display a type of symmetry, and in particular, if $G \xrightarrow[\{i\}]{} G'$, then $G' \xrightarrow[\{i\}]{} G$.

Under noncooperative rules, the effectiveness relations over networks G_1 and G_2 in Figure 2 are given by

$$G_1 \xrightarrow[\{i_3\}]{} G_2 \quad G_2 \xrightarrow[\{i_3\}]{} G_1.$$

Note that under noncooperative rules, networks G_0 and G_3 in Figure 3 are *not* related under the effectiveness relations $\{\rightarrow_{\{i\}}\}_{i \in N}$. However, referring to the networks in Figures 2 and 3, under the noncooperative rules we have, for example, the following effectiveness relations

$G_3 \rightarrow_{\{i_2\}} G_2 \quad G_2 \rightarrow_{\{i_3\}} G_0$
and
$G_0 \rightarrow_{\{i_3\}} G_2 \quad G_2 \rightarrow_{\{i_2\}} G_3$.

(½,¾)-*voting Rules*: All of the rules above require that arc addition and arc subtraction involve at least one player who is a party to the arc. Consider now arc addition and arc subtraction based on voting. If nodes represent players, then under certain voting rules, arcs can be imposed on players. To see this, consider the following rules for arc addition and arc subtraction.

(i) adding an arc a from player i to player i' requires a simple majority agree to add arc a;
(ii) subtracting an arc a from player i to player i' requires a ¾ majority agree to subtract arc a;
(iii) for any pair of networks G and G' in $\mathbb{G}$, if $G \rightarrow_S G'$, then $G \neq G'$ and

either $G' = G \cup (a, (i, i'))$ for some $(a, (i, i')) \in A \times (N \times N)$
or
$G' = G \backslash (a, (i, i'))$ for some $(a, (i, i')) \in A \times (N \times N)$

(i.e., networks changes take place one arc at a time).

We shall refer to rules (i)–(iii) above as (½,¾)-voting rules. Thus, under rules (i)–(iii), if $G \rightarrow_S G'$, then $G \neq G'$ and either

$G' = G \cup (a, (i, i'))$ for some $(a, (i, i')) \in A \times (N \times N)$ and $\frac{|S|}{|D|} \geq \frac{1}{2}$
or
$G' = G \backslash (a, (i, i'))$ for some $(a, (i, i')) \in A \times (N \times N)$ and $\frac{|S|}{|D|} \geq \frac{3}{4}$

Referring to Figure 2, the effectiveness relations over networks G_1 and G_2 under (½,¾)-voting rules are given by

$$G_1 \xrightarrow[\{i_1,i_2\}]{} G_2 \quad G_1 \xrightarrow[\{i_1,i_3\}]{} G_2 \quad G_1 \xrightarrow[\{i_2,i_3\}]{} G_1 \quad G_1 \xrightarrow[\{i_1,i_2,i_3\}]{} G_2$$

and

$$G_2 \xrightarrow[\{i_1,i_2,i_3\}]{} G_1.$$

Note that under (½,¾)-voting rules the move from network G_1 to network G_2 may involve the imposition of arc a_1 from player i_3 to player i_1 upon player i_1 by players i_2 and players i_3. Also, note that under (½,¾)-voting rules in order to move from network G_2 back to network G_1 (i.e., in order to remove arc a_1 from player i_3 to player i_1) requires the agreement of all three players.

Nonuniform Rules and the Network Representation of Network Formation Rules. In all of the examples above, the rules for arc addition and arc subtraction are uniform across pairs of networks. In some applications, such uniformity is not present. One very concise way to write down such nonuniform network formation rules is to use a network representation. In particular, suppose we write

$(S,(G,\ G'))$ if and only if $G \rightarrow_S G'$.

Thus, $(S,(G, G'))$ if and only if under the rules coalition $S \in P(D)$ can change network G to network G'. Letting the set of arcs be given by the collection of all coalitions $P(D)$ and letting the set of nodes be given by the feasible set of networks $\mathbb{G}$, the rules of network formation can be represented by a network $\mathbf{G} \subseteq P(D) \times (\mathbb{G} \times \mathbb{G})$. Then the set of all possible network formation rules is given by the set of all such networks.

3.3 Dominance relations

We will focus primarily on two types of dominance relations on the feasible set of networks $\mathbb{G} \subseteq P(A \times (N \times N))$, direct and indirect dominance.

3.3.1 Direct Dominance

Network $G' \in \mathbb{G}$ *directly dominates* network $G \in \mathbb{G}$, sometimes written $G' \triangleright G$, if for some coalition $S \in P(D)$,

$$G \prec_S G'$$

and

$$G \xrightarrow[S]{} G'.$$

Thus, network G' directly dominates network G if some coalition S prefers G' to

G and if under the rules of network formation coalition S has the power to change G to G'.

3.3.2 Indirect dominance

Network $G' \in \mathbb{G}$ *indirectly dominates* network $G \in \mathbb{G}$, written $G' \rhd\rhd G$, if there is a *finite* sequence of networks,

$$G_0, G_1, \ldots, G_h,$$

with $G = G_0$, $G' = G_h$, and $G_k \in \mathbb{G}$ for $k = 0, 1, \ldots, h$, and a corresponding sequence of coalitions,

$$S_1, S_2, \ldots, S_h,$$

such that for $k = 1, 2, \ldots, h$

$$G_{k-1} \xrightarrow[S_k]{} G_k,$$

and

$$G_{k-1} \prec_{S_k} G_h.$$

Note that if network G' indirectly dominates network G (i.e., if $G' \rhd\rhd G$), then what matters to the initially deviating coalition S_1, as well as all the coalitions along the way, is that the ultimate network outcome $G' = G_h$ be preferred. Thus, for example, the initially deviating coalition S_1 will not be deterred from changing network G_0 to network G_1 even if network G_1 is not preferred to network $G = G_0$, as long as the ultimate network outcome $G' = G_h$ is preferred to G_0, that is, as long as $G_0 \prec_{S_1} G_h$.[20]

3.3.3 Path dominance

Each dominance relation $>$ induces a path dominance relation on the set of networks. In particular, corresponding to dominance relation $>$ on networks $\mathbb{G}$ there is a corresponding path dominance relation $\geq_p$ on $\mathbb{G}$ specified as follows: network $G' \in \mathbb{G}$ (weakly) path dominates network $G \in \mathbb{G}$ with respect to $>$ (i.e., with respect to the underlying dominance relation $>$), written $G' \geq_p G$, if $G' = G$ or if there exists a *finite* sequence of networks $\{G_k\}_{k=0}^{h}$ in $\mathbb{G}$ with $G_h = G'$ and $G_0 = G$ such that for $k = 1, 2, \ldots, h$

$$G_k > G_{k-1}.$$

We refer to such a finite sequence of networks as a *finite domination path* and we say

[20] In order to capture the idea of farsightedness in strategic behavior, Chwe (1994) analyzes abstract games equipped with indirect dominance relations in great detail, introducing the equilibrium notions of consistency and largest consistent set. The basic idea of indirect dominance goes back to the work of Guilbaud (1949) and Harsanyi (1974).

network G' is >-*reachable* from network G if there exists a finite domination path from G to G'. Thus,

$$G' \geq_p G \text{ if and only if } \begin{cases} G' \text{ is} > -\text{recheable from } G, \text{ or} \\ G' = G \end{cases} \tag{2}$$

If network G is reachable from network G, that is, if there is a finite domination path from G back to G then we call this path a *circuit*. Finally, if network G is *not* reachable from any network in $\mathbb{G}$ and if no network in $\mathbb{G}$ is reachable from G, then network G is *isolated* (i.e., network $G \in \mathbb{G}$ is isolated if there does not exist a network $G' \in \mathbb{G}$ with $G' \geq_p G$ or $G \geq_p G'$).

3.3.4 The directed graph of a dominance relation

It is often useful to represent the dominance relation over networks using a directed graph. For example, Figure 4 depicts the graph of dominance relation > on the feasible set of networks $\mathbb{G} = \{G_0, G_1, \ldots, G_7\}$.

The arrow (or >-arc) *from* network G_3 *to* network G_4 in Figure 4 indicates that G_4 dominates G_3. Given primitives $(\mathbb{G}, \{ \succ_S\}, \{\rightarrow_S\}, >)_{S \in P(D)}$ and given that > is a *direct* dominance relation, the >-arc from network G_3 to network G_4 means that for some coalition S, G_4 is preferred to G_3 and more importantly, that coalition S has the power to change network G_3 to network G_4. Thus, $G_3 \prec_S G_4$ and $G_3 \rightarrow_S G_4$. But notice also that there is a >-arc in the opposite direction, from network G_4 to network G_3. Thus, G_3 also dominates G_4, and thus for some other coalition S'

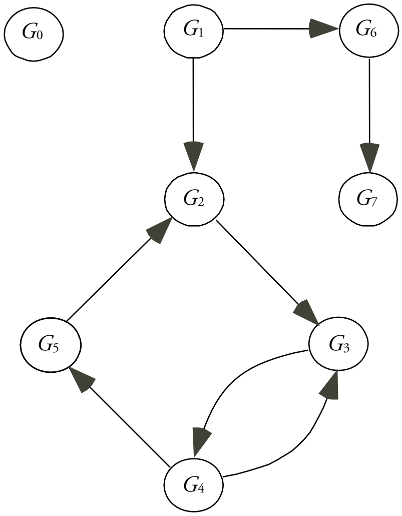


Figure 4: Directed Graph of Dominance Relation >

distinct from coalition S, that is, some coalition S' with $S' \cap S = \varnothing$, $G_4 \prec_{S'} G_3$ and $G_4 \rightarrow_{S'} G_3$.[21]

Note that network G_3 is >-reachable from network G_3 via two paths. Thus, the graph of dominance relation > depicted in Figure 4 contains two circuits. Defining the *length* of a domination path to be the number of >-arcs in the path, these two circuits are of length 4 and length 2.

Because networks G_2 and G_5 in Figure 4 are on the same circuit, G_5 is >-reachable from G_2 *and* G_2 is >-reachable from G_5. Thus, G_5 path dominates G_2 (i.e., $G_5 \geq_p G_2$) *and* G_2 path dominates G_5 (i.e., $G_2 \geq_p G_5$). The same cannot be said of networks G_1 and G_5 in Figure 4. In particular, while $G_5 \geq_p G_1$, it is not true that $G_1 \geq_p G_5$ because G_1 is not >-reachable from G_5. Finally, note that network G_0 is isolated.

4. Network Formation Games and Stability

We can now present our main results. Using the abstract network formation game with respect to path dominance given by the pair

$$(\mathbb{G}, \geq_p) \tag{3}$$

and induced by primitives

$$(\mathbb{G}, \{\succ_S\},\{\rightarrow_S\}, >)_{S \in P(D)}, \tag{4}$$

we introduce and characterize the notions of (i) strategic basins of attraction, (ii) path dominance stable sets, and (iii) the path dominance core. All of the results presented in this section hold for any path dominance relation induced by an irreflexive dominance relation constructed from coalitional preferences, $\{\succ_S\}_{S \in P(D)}$ and coalitional effectiveness relations, $\{\rightarrow_S\}_{S \in P(D)}$.[22]

4.1 Networks without descendants

If $G_1 \geq_p G_0$ and $G_0 \geq_p G_1$, networks G_1 and G_0 are *equivalent*, written $G_1 \equiv_p G_0$. If networks G_1 and G_0 are equivalent then either networks G_1 and G_0 coincide or G_1 and G_0 are on the same circuit (see Figure 4 above for a picture of a circuit). If $G_1 \geq_p G_0$ but G_1 and G_0 are not equivalent (i.e., not $G_1 \equiv_p \mathrm{G}_0$), then network G_1 is a *descendant* of network G_0 and we write

$$G_1 >_p G_0. \tag{5}$$

[21] Note that if preferences over networks are weak as in Remark 1, then the statement, *for some other coalition S' distinct from coalition S* can be weakened to *for some other coalition S' not equal to coalition S.* With this weakening, the requirement that the intersection of S and S' be empty is no longer required.

[22] In fact, all the results in this section hold for any abstract game $(\mathbb{G}, \geq_p)$ where G is a finite set of outcomes and $\geq_p$ is a path dominance relation induced by any binary relation on $\mathbb{G}$.

Referring to Figure 4, observe that network G_5 is a descendant of network G_1, that is, $G_5 >_p G_1$.

Network $G' \in \mathbb{G}$ *has no descendants in* $\mathbb{G}$ if for any network $G \in \mathbb{G}$

$$G \geq_p G' \text{ implies that } G \equiv_p G'.$$

Thus, if G' has no descendants then $G \geq_p G'$ implies that G and G' coincide or lie on the same circuit.[23]

In attempting to identify those networks which are likely to emerge and persist, networks *without descendants* are of particular interest. Here is our main result concerning networks without descendants.

Theorem 1 – All path dominance network formation games have networks without descendants. *Let $(\mathbb{G}, \geq_p)$ be a network formation game. For every network $G \in \mathbb{G}$ there exists a network $G' \in \mathbb{G}$ such that $G' \geq_p G$ and G' has no descendants.*

Proof. Let G_0 be any network in $\mathbb{G}$. If G_0 has no descendants then we are done. If not choose G_1 such that $G_1 >_p G_0$. If G_1 has no descendants then we are done. If not, continue by choosing $G_2 >_p G_1$. Proceeding iteratively, we can generate a sequence, $G_0, G_1, G_2, \ldots$. Now observe that in a finite number of iterations we must come to a network $G_{k'}$ without descendants. Otherwise, we could generate an infinite sequence, $\{G_k\}_k$ such that for all k,

$$G_k >_p G_{k-1}.$$

However, because $\mathbb{G}$ is finite this sequence would contain at least one network, say $G_{k'}$, which is repeated an infinite number of times. Thus, all the networks in the sequence lying between any two consecutive repetitions of $G_{k'}$ would be on the same circuit, contradicting the fact that for all k, G_k is a descendant of G_{k-1} (i.e., $G_k >_p G_{k-1}$). ■

By Theorem 1, in any network formation game $(\mathbb{G}, \geq_p)$, corresponding to any network $G \in \mathbb{G}$ there is a network $G' \in G$ without descendants which is >-reachable from G. Thus, in any network formation game the set of networks without descendants is nonempty. Referring to Figure 4, the set of networks without descendants is given by

$$\{G_0, G_2, G_3, G_4, G_5, G_7\}.$$

We shall denote by $\mathbb{Z}$ the set of networks without descendants.

[23] Note that any isolated network is by definition a network without descendants (e.g., network G_0 in Figure 3).

4.2 Basins of attraction

Stated loosely, a basin of attraction is a set of *equivalent* networks to which the strategic network formation process represented by the game might tend and from which there is no escape. Formally, we have the following definition.

Definition 2 – Basin of Attraction. Let $(\mathbb{G}, \geq_p)$ be a network formation game. A set of networks $\mathbb{A} \subseteq \mathbb{G}$ is said to be a basin of attraction for $(\mathbb{G}, \geq_p)$ if

1. the networks contained in $\mathbb{A}$ are equivalent (i.e., for all G' and G in $\mathbb{A}$, $G' \equiv_p G$) and for no set $\mathbb{A}'$ having $\mathbb{A}$ as a strict subset is this true that all the networks in $\mathbb{A}'$ are equivalent,[24] and
2. no network in $\mathbb{A}$ has descendants (i.e., there does not exist a network $G' \in \mathbb{G}$ such that $G' >_p \mathbb{G}$ for some $G \in \mathbb{A}$).

As the following characterization result shows, there is a very close connection between networks without descendants and basins of attraction.

Theorem 2 – A characterization of basins of attraction. *Let* $(\mathbb{G}, \geq_p)$ *be a network formation game and let* $\mathbb{A}$ *be a subset of networks in* $\mathbb{G}$*. The following statements are equivalent:*

1. $\mathbb{A}$ *is a basin of attraction for* $(\mathbb{G}, \geq_p)$.
2. *There exists a network without descendants,* $G \in \mathbb{Z}$*, such that*

$$\mathbb{A} = \{G' \in \mathbb{Z}: G' \equiv_p G\} .$$

Proof. (1) implies (2): Because the sets $\mathbb{A}$ and $\{G' \in \mathbb{Z}: G' \equiv_p G\}$, $G \in \mathbb{Z}$, are equivalence classes, $\mathbb{A} \neq \{G' \in \mathbb{Z}: G' \equiv_p G\}$ implies that

$$\mathbb{A} \cap \{G' \in \mathbb{Z}: G' \equiv_p G\} = \varnothing \text{ for all } G \in \mathbb{Z}.$$

Thus, if (2) fails, this implies that $\mathbb{A}$ contains a network with descendants. Thus, $\mathbb{A}$ cannot be a basin of attraction for $(\mathbb{G}, \geq_p)$, and thus, (1) implies (2).[25]

(2) implies (1): Suppose now that

$$\mathbb{A} = \{G' \in \mathbb{Z}: G' \equiv_p G\}$$

for some network $G \in \mathbb{Z}$. If $\mathbb{A}$ is not a basin of attraction, then for some network $G'' \in G$, $G'' >_p G'$ for some $G' \in \mathbb{A}$. But now $G'' >_p G'$and $G' \equiv_p G$ imply that $G'' >_p G$, contradicting the fact that $G \in \mathbb{Z}$. Thus, (2) implies (1). ■

[24] $\mathbb{A}$ is a strict subset of $\mathbb{A}'$ if

$$\mathbb{A} \subset \mathbb{A}' \text{ and } \mathbb{A}' \backslash \mathbb{A} \neq \varnothing.$$

[25] Note that if $G \in \mathbb{Z}$ and $G' \equiv_p$ G, then $G' \in \mathbb{Z}$.

In light of Theorem 2, we conclude that in any network formation game $(\mathbb{G}, \geq_p)$, $\mathbb{G}$ contains a *unique*, finite, disjoint collection of basins of attraction, say $\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\}$, where for each $k = 1, 2, \ldots, m$ $(m \geq 1)$

$$\mathbb{A}_k = \mathbb{A}_G := \{G' \in \mathbb{Z}: \ G' \equiv_p G\}$$

for some network $G \in \mathbb{Z}$. Note that for networks G' and G in $\mathbb{Z}$ such that $G' \equiv_p G$, $\mathbb{A}_{G'} = \mathbb{A}_G$ (i.e. the basins of attraction $\mathbb{A}_{G'}$ and $\mathbb{A}_G$ coincide). Also, note that if network $G \in \mathbb{G}$ is isolated, then $G \in \mathbb{Z}$ and

$$\mathbb{A}_G := \{G' \in \mathbb{Z}: \ G' \equiv_p G\} = \{G\}$$

is, by definition, a basin of attraction – but a very uninteresting one.

EXAMPLE 1 – *Basins of attraction*. In Figure 4 above the set of networks without descendants is given by

$$\mathbb{Z} = \{G_0, G_2, G_3, G_4, G_5, G_7\}.$$

Even though there are six networks without descendants, because networks G_2, G_3, G_4, and G_5 are equivalent, there are only three basins of attraction:

$$\mathbb{A}_1 = \{G_0\}, \quad \mathbb{A}_2 = \{G_2, G_3, G_4, G_5\}, \quad \text{and} \quad \mathbb{A}_3 = \{G_7\}.$$

Moreover, because G_2, G_3, G_4, and G_5 are equivalent,

$$\mathbb{A}_{G_2} = \mathbb{A}_{G_3} = \mathbb{A}_{G_4} = \mathbb{A}_{G_5} = \{G_2, G_3, G_4, G_5\}.$$

4.3 Stable sets with respect to path dominance

The formal definition of a $\geq_p$-stable set is as follows.[26]

Definition 3 – Stable sets with respect to path dominance. Let $(\mathbb{G}, \geq_p)$ be a network formation game. A subset $\mathbb{V}$ of networks in $\mathbb{G}$ is said to be a stable set for $(\mathbb{G}, \geq_p)$ if

(a) (internal $\geq_p$-stability) whenever G_0 and G_1 are in $\mathbb{V}$, with $G_0 \neq G_1$, then neither $G_1 \geq_p G_0$ nor $G_0 \geq_p G_1$ hold, and

(b) (external $\geq_p$-stability) for any $G_0 \notin \mathbb{V}$ there exists $G_1 \in \mathbb{V}$ such that $G_1 \geq_p G_0$.

In other words, a nonempty subset of networks $\mathbb{V}$ is a stable set for $(\mathbb{G}, \geq_p)$ if G_0 and G_1 are in $\mathbb{V}$, with $G_0 \neq G_1$, then G_1 is not reachable from G_0, nor is G_0 reachable from G_1, and if $G_0 \notin \mathbb{V}$, then there exists $G_1 \in \mathbb{V}$ reachable from G_0.

[26] By equipping the abstract network formation game with the path dominance relation rather than the original dominance relation, we entirely avoid the famous Lucas (1968) example of a game with no stable set.

We now have our main results on the existence, construction, and cardinality of stable sets.[27]

Theorem 3 – Stable sets: Existence, construction, and cardinality. *Let* $(\mathbb{G}, \geq_p)$ *be a network formation game, and without loss of generality assume that* $(\mathbb{G}, \geq_p)$ *has basins of attraction given by*

$$\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\},$$

where basin of attraction $\mathbb{A}_k$ *contains* $|\mathbb{A}_k|$ *many networks (i.e.,* $|\mathbb{A}_k|$ *is the cardinality of* $\mathbb{A}_k$*). Then the following statements are true:*

1. $\mathbb{V} \subseteq \mathbb{G}$ *is a stable set for* $(\mathbb{G}, \geq_p)$ *if and only if* $\mathbb{V}$ *is constructed by choosing one network from each basin of attraction, that is, if and only if* $\mathbb{V}$ *is of the form*

   $$\mathbb{V} = \{G_1, G_2, \ldots, G_m\},$$

   where $G_k \in \mathbb{A}_k$ *for* $k = 1, 2, \ldots, m$.
2. $(\mathbb{G}, \geq_p)$ *possesses*

   $$|\mathbb{A}_1| \cdot |\mathbb{A}_2| \cdot\cdots\cdot |\mathbb{A}_m| := M$$

   many stable sets and each stable set, $\mathbb{V}_q$*,* $q = 1, 2, \ldots, M$*, has cardinality*

   $$|\mathbb{V}_q| = |\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\}| = m.$$

Proof. It suffices to prove (1). Given (1), the proof of (2) is straightforward. To begin, let

$$\mathbb{V} = \{G_1, G_2, \ldots, G_m\},$$

where $G_k \in \mathbb{A}_k$ for $k = 1, 2, \ldots, m$, and suppose that for G_k and $G_{k'}$ in $\mathbb{V}$, $G_{k'} \geq_p G_k$. Since $G_k \in \mathbb{A}_k$ has no descendants, this would imply that $G_{k'} \equiv_p G_k$. But this is a contradiction because $G_k \in \mathbb{A}_k$ and $G_{k'} \in \mathbb{A}_{k'}$ and the basins of attraction $\mathbb{A}_k$ and $\mathbb{A}_{k'}$ are disjoint. Thus, $\mathbb{V}$ is internally $\geq_p$-stable. Now suppose that network G is not contained in $\mathbb{V}$. By Theorem 1, there exists a network $G' \in \mathbb{G}$ without descendants such that $G' \geq_p G$. By Theorem 2, G' is contained in some basin of attraction $\mathbb{A}_k$ and therefore $G' \equiv_p G_k$ where G_k is the kth component of $\{G_1, G_2, \ldots, G_m\}$. Thus, we have $G_k \geq_p G' \geq_p G$ implying that $G_k \geq_p G$, and thus $\mathbb{V}$ is externally $\geq_p$-stable.

Suppose now that $\mathbb{V} \subseteq \mathbb{G}$ is a stable set for $(\mathbb{G}, \geq_p)$. First note that each network G in $\mathbb{V}$ is a network without descendants. Otherwise there exists $G' \in \mathbb{G}\backslash\mathbb{V}$ such that $G' >_p G$. But then because $\mathbb{V}$ is externally $\geq_p$-stable, there exists $G'' \in \mathbb{V}$,

[27] These results can be viewed as applications of some classical results from graph theory to the theory of network formation games (e.g., see Berge 2001, Chapter 2).

$G'' \neq G$, such that $G'' \geq_p G'$ implying that $G'' \geq_p G$ and contradicting the internal $\geq_p$-stability of $\mathbb{V}$. Because each $G \in \mathbb{V}$ is without descendants, it follows from Theorem 2 that each $G \in \mathbb{V}$ is contained in some basin of attraction $\mathbb{A}_k$. Moreover, because $\mathbb{V}$ is internally $\geq_p$-stable and because all networks contained in any one basin of attraction are equivalent, no two distinct networks contained in $\mathbb{V}$ can be contained in the same basin of attraction. It only remains to show that for each basin of attraction, $\mathbb{A}_k$, $k = 1, 2, \ldots, m$,

$$\mathbb{V} \cap \mathbb{A}_k \neq \varnothing.$$

Suppose not. Then for some k', $\mathbb{V} \cap \mathbb{A}_{k'} = \varnothing$. Because all networks in $\mathbb{A}_{k'}$ are without descendants, for no network $G \in \mathbb{A}_{k'}$ is it true that there exists a network $G' \in \mathbb{V}$ such that $G' \geq_p G$. Thus, we have a contradiction of the external $\geq_p$-stability of $\mathbb{V}$. ■

EXAMPLE 2 – *Basins of attraction and stable sets*. Referring to Figure 4, it follows from Theorem 3 that because

$$|\mathbb{A}_1| \cdot |\mathbb{A}_2| \cdot |\mathbb{A}_3| = 1 \cdot 4 \cdot 1 = 4,$$

the network formation game $(\mathbb{G}, \geq_p)$ has 4 stable sets, each with cardinality 3. By examining Figure 4 in light of Theorem 3, we see that the stable sets for $(\mathbb{G}, \geq_p)$ are given by

$$\mathbb{V}_1 = \{G_0, G_2, G_7\},$$
$$\mathbb{V}_2 = \{G_0, G_3, G_7\},$$
$$\mathbb{V}_3 = \{G_0, G_4, G_7\},$$
$$\mathbb{V}_4 = \{G_0, G_5, G_7\}.$$

4.4 The path dominance core

Definition 4 – The path dominance core. Let $(\mathbb{G}, \geq_p)$ be a network formation game. A network $G \in \mathbb{G}$ is contained in the path dominance core $\mathbb{C} \subset \mathbb{G}$ if and only if there does not exist a network $G' \in \mathbb{G}$, $G' \neq G$, such that $G' \geq_p G$.

Our next results give necessary and sufficient conditions for the path dominance core of a network formation game to be nonempty, as well as a recipe for constructing the path dominance core.

Theorem 4 – Path dominance core: Nonemptiness and construction. *Let $(\mathbb{G}, \geq_p)$ be a network formation game, and without loss of generality assume that $(\mathbb{G}, \geq_p)$ has basins of attraction given by*

$\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\}$,

where basin of attraction $\mathbb{A}_k$ contains $|\mathbb{A}_k|$ many networks. Then the following statements are true:

1. *$(\mathbb{G}, \geq_p)$ has a nonempty path dominance core if and only if there exists a basin of attraction containing a single network, that is, if and only if for some basin of attraction $\mathbb{A}_k$, $|\mathbb{A}_k| = 1$.*
2. *Let*

   $\{\mathbb{A}_{k_1}, \mathbb{A}_{k_2}, \ldots, \mathbb{A}_{k_n}\} \subseteq \{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\}$

   be the subset of basins of attraction containing all basins having cardinality 1. Then the path dominance core $\mathbb{C}$ of $(\mathbb{G}, \geq_p)$ is given by

   $\mathbb{C} = \{G_{k_1}, G_{k_2}, \ldots, G_{k_n}\}$,

   where $G_{k_i} \in \mathbb{A}_{k_i}$, for $i = 1, 2, \ldots, n$.

Proof. It suffices to show that a network G is contained in the path dominance core $\mathbb{C}$ if and only if $G \in \mathbb{A}_k$ for some basin of attraction $\mathbb{A}_k$, $k = 1, 2, \ldots, m$, with $|\mathbb{A}_k| = 1$. First note that if G is in the path dominance core, then G is a network without descendants. Thus, $G \in \mathbb{A}_k$ for some basin of attraction $\mathbb{A}_k$. If $|\mathbb{A}_k| > 1$, then there exists another network $G' \in \mathbb{A}_k$ such that $G' \equiv_p G$. Thus, $G' \geq_p G$ contradicting the fact that G is in the path dominance core. Conversely, if $G \in \mathbb{A}_k$ for some basin of attraction $\mathbb{A}_k$ with $|\mathbb{A}_k| = 1$, then there does not exist a network $G' \neq G$ such that $G' \geq_p G$. ■

Remark 2. If coalitional preferences $\{\succ_S\}_{S \in P(D)}$ over networks are based on weak preference relations $\{\succsim_d\}_{d \in D}$ rather than on strong preference relations $\{\succ_d\}_{d \in D}$ (see Remark 1 above), then the corresponding path dominance core – the weak path dominance core – is contained in the path dominance core based on strong preference relations.

EXAMPLE 3 – *Basins of attraction and the path dominance core.* It follows from Theorem 4 that the path dominance core of the network formation game $(\mathbb{G}, \geq_p)$ with feasible set

$\mathbb{G} = \{G_0, G_1, \ldots, G_7\}$

and path dominance relation $\geq_p$ induced by the dominance relation depicted in Figure 4 is

$\mathbb{C} = \{G_0, G_7\}$.

Figure 5 contains the graph of a different dominance relation on $\mathbb{G} = \{G_0, G_1, \ldots, G_7\}$.

Denoting the new dominance relation by $>$, the network formation game $(\mathbb{G}, \geq_p)$ with respect to the path dominance relation $\geq_p$ induced by the dominance relation $>$has 3 circuits and 2 basins of attraction,

$\mathbb{A}_1 = \{G_2, G_3, G_4, G_5\}$ and $\mathbb{A}_2 = \{G_6, G_7\}$.

Because $|\mathbb{A}_1| = 4$ and $|\mathbb{A}_2| = 2$, by Theorem 4 the path dominance core of $(\mathbb{G}, \geq_p)$ is empty. By Theorem 3, $(\mathbb{G}, \geq_p)$ has 8 stable sets each containing 2 networks (i.e., each with cardinality 2). These stable sets are given by

$\mathbb{V}_1 = \{G_2, G_6\}$,
$\mathbb{V}_2 = \{G_3, G_6\}$,
$\mathbb{V}_3 = \{G_4, G_6\}$,
$\mathbb{V}_4 = \{G_5, G_6\}$,
$\mathbb{V}_5 = \{G_2, G_7\}$,
$\mathbb{V}_6 = \{G_3, G_7\}$,
$\mathbb{V}_7 = \{G_4, G_7\}$,
$\mathbb{V}_8 = \{G_5, G_7\}$.

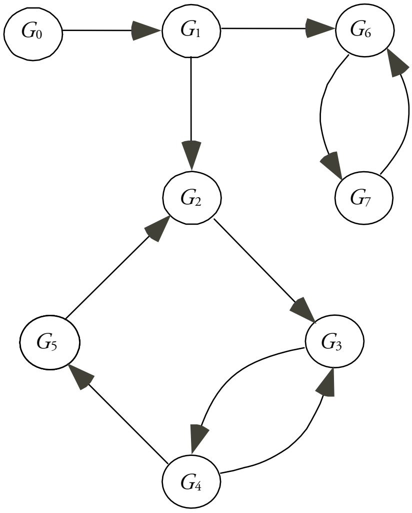


Figure 5: Graph of a different dominance relation >

4.4.1 The path dominance core and constrained Pareto efficiency

Given primitives $(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S \in P(D)}$, we say that a network $G \in \mathbb{G}$ is constrained Pareto efficient if and only if there does not exist another network $G' \in \mathbb{G}$ such that (i) some coalition S can change network G to network G' (that is, $G \rightarrow_S G'$ for some coalition $S \in P(D)$) and (ii) G' is preferred by all players (that is, $G \prec^{d} G'$

for *all* players $d \in D$). Letting $\mathbb{E}$ denote the set of all constrained Pareto efficient networks, it is easy to see that the path dominance core $\mathbb{C}$ of network formation game $(\mathbb{G}, \geq_p)$ is a subset of $\mathbb{E}$, that is, $\mathbb{C} \subseteq \mathbb{E}$.

Under the classical notion of Pareto efficiency, a network G is said to be Pareto efficient if and only if there does not exists another network G' such that $G \prec_d G'$ for *all* players $d \in D$, regardless of whether or not some coalition S can change network G to network G'. Letting $\mathbb{PE}$ denote the set of all classically Pareto efficient networks, it is easy to see that $\mathbb{PE} \subseteq \mathbb{E}$. Note, however, that if under primitives $(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S \in P(D)}$, any network G can be changed to any other network G' via the actions of some coalition S, then the notions of constrained Pareto efficiency and classical Pareto efficiency are equivalent. Thus, if the collection of coalitional effectiveness relations $\{\rightarrow_S\}_{S \in P(D)}$ on $\mathbb{G}$ is complete, that is, if for any pair of networks G and G' in $\mathbb{G}$, $G \rightarrow_S G'$ for some coalition $S \in P(D)$, then $\mathbb{PE} = \mathbb{E}$, and we have $\mathbb{C} \subseteq \mathbb{PE} = \mathbb{E}$.

5. Other Stability Notions for Network Formation Games

5.1 Strongly stable networks

We begin with a formal definition of strong stability for abstract network formation games.

Definition 5 – Strong stability. Given primitives $(\mathbb{G},\{\succ_S\},\{\rightarrow_S\},>)_{S \in P(D)}$ and network formation game $(\mathbb{G}, \geq_p)$, network $G \in \mathbb{G}$ is said to be strongly stable in $(\mathbb{G}, \geq_p)$ if for all $G' \in \mathbb{G}$ and $S \in P(D)$, $G \rightarrow_S G'$ implies that $G \not\prec_S G'$.

Thus, a network is strongly stable if whenever a coalition has the power to change the network to another network, the coalition will be deterred from doing so because not *all* members of the coalition are made better off by such a change.[28] If nodes represent players and arc addition is bilateral while arc subtraction is unilateral, then our definition of strong stability is essentially that of Jackson-van den Nouweland but for directed networks rather than linking networks. Note that under our definition of strong stability a network $G \in \mathbb{G}$ that cannot be changed to another network by any coalition is strongly stable.

We now have our main result on the path dominance core and strong stability. Denote the set of strongly stable networks by $\mathbb{SS}$.

[28] Our definition of a strongly stable network differs slightly from the definition given in Jackson-van den Nouweland (2005). In particular, under their definition, a network is strongly stable if whenever a coalition has the power to change the network to another network, the coalition will be deterred from doing so because at least one member of the coalition is made *worse off* by the change. If coalitional preferences, $\{\succ_S\}_{S \in P(D)}$ are based upon weak players preferences, $\{\succsim_d\}_{d \in D}$, then our definition of strong stability is equivalent to that of Jackson-van den Nouweland (see Remark 1). As it stands, our definition is closely related to that given by Dutta and Mutuswami (1997).

Theorem 5 – The path dominance core and strong stability*. Given primitives* $(\mathbb{G}, \{\succ_S\}, \{\to_S\}, >)_{S \in P(D)}$ *and network formation game* $(\mathbb{G}, \geq_p)$, *where path dominance* $\geq_p$ *is induced by either a direct relation or an indirect dominance relation, the following statements are true.*

1. *If the path dominance core* $\mathbb{C}$ *of* $(\mathbb{G}, \geq_p)$ *is nonempty, then* $\mathbb{SS}$ *is nonempty and* $\mathbb{C} \subseteq \mathbb{SS}$.
2. *If the dominance relation > underlying* $\geq_p$ *is a direct dominance relation, then* $\mathbb{C} = \mathbb{SS}$ *and* $\mathbb{SS}$ *is nonempty if and only if there exists a basin of attraction containing a single network.*

Proof. Let $\mathbb{C} \subseteq \mathbb{G}$, $\mathbb{C} \neq \varnothing$, be the path dominance core of $(\mathbb{G}, \geq_p)$ and let network G be contained in $\mathbb{C}$. Then there does not exist a network $G' \in \mathbb{G}$, $G' \neq G$, such that $G' \geq_p G$. If for some coalition S and some network $G' \in \mathbb{G}$, $G \to_S G'$ *and* $G \prec_S G'$, then $G' \geq_p G$ trivially, a contradiction. Thus, for G contained in $\mathbb{C}$, $G \to_S G'$ for coalition S implies that $G \nprec_S G'$, and thus $G \in \mathbb{C}$ implies $G \in \mathbb{SS}$.

2. To see that $\mathbb{SS} \subseteq \mathbb{C}$ if the dominance relation > underlying $\geq_p$ is a direct dominance relation, consider the following. If $G \notin \mathbb{C}$, then there exists a network $G' \neq G$ which path dominates G, that is, $G' \geq_p G$. This implies that there exists a network G'' such that $G' \geq_p G'' > G$. Because > is a direct dominance relation, for some coalition S we have $G \to_S G''$ and $G \prec_S G''$. Thus, $G \notin \mathbb{SS}$. By part 1 of Theorem 4, $\mathbb{C} = \mathbb{SS}$ is nonempty if and only if there exists a basin of attraction containing a single network. ■

Note that the set of strongly stable networks is contained in the set of constrained Pareto efficient networks. Thus, $\mathbb{C} \subseteq \mathbb{SS} \subseteq \mathbb{E}$.

5.2 Pairwise stable networks

The following definition is a formalization of Jackson-Wolinsky (1996) pairwise stability for abstract network formation games.

Definition 6 – Pairwise stability. Given networks $P(A \times (N \times N))$ where nodes represent players (i.e., $N = D$) and given feasible networks $\mathbb{G} \subseteq P(A \times (N \times N))$ and primitives $(\mathbb{G}, \{\succ_S\}, \{\to_S\}, >)_{S \in P(D)}$, network $G \in \mathbb{G}$ is said to be pairwise stable in network formation game $(\mathbb{G}, \geq_p)$ if for all $(a, (i, i')) \in A \times (N \times N)$,

1. $G \to_{\{i,i'\}} G \cup (a, (i, i'))$ implies that $G \nprec_{\{i,i'\}} G \cup (a, (i, i'))$;
2. (a) $G \to_{\{i\}} G \backslash (a, (i, i'))$ implies that $G \nprec_{\{i\}} G \backslash (a, (i, i'))$, and
   (b) $G \to_{\{i'\}} G \backslash (a, (i, i'))$ implies $G \nprec_{\{i'\}} G \backslash (a, (i, i'))$.

Thus, a network is pairwise stable if there is no incentive for any pair of

(b) $G \rightarrow_{\{i'\}} G \setminus (a, (i, i'))$ implies $G \not\prec_{\{i'\}} G \setminus (a, (i, i'))$.

Thus, a network is pairwise stable if there is no incentive for any pair of players to add an arc to the existing network and there is no incentive for any player who is party to an arc in the existing network to dissolve or remove the arc. Note that under our definition of pairwise stability a network $G \in \mathbb{G}$ that cannot be changed to another network by any coalition, or can only be changed by coalitions of size greater than 2, is pairwise stable.

Let $\mathbb{PS}$ denote the set of pairwise stable networks. It follows from the definitions of strong stability and pairwise stability that $\mathbb{SS} \subseteq \mathbb{PS}$. Moreover, if the full set of Jackson-Wolinsky rules are in force, then $\mathbb{SS} = \mathbb{PS}$. Jackson-van den Nouweland (2005) provide two examples of the potential for strong stability to refine pairwise stability (i.e., two examples where $\mathbb{SS}$ is a strict subset of $\mathbb{PS}$). However, *under Jackson-Wolinsky rules* because network changes can occur only one arc at a time and because deviations by coalitions of more than two players are not possible such refinements are not possible driving $\mathbb{SS}$ and $\mathbb{PS}$ to equality.[29]

We now have our main result on the path dominance core and pairwise stability.

Theorem 6 – The path dominance core and pairwise stability. *Given primitives* $(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S \in P(D)}$ *where nodes represent players (i.e., $N = D$) and given network formation game* $(\mathbb{G}, \geq_p)$, *where path dominance $\geq_p$ is induced by either a direct relation or an indirect dominance relation, the following statements are true.*

1. *If the path dominance core $\mathbb{C}$ of $(\mathbb{G}, \geq_p)$ is nonempty, then $\mathbb{PS}$ is nonempty and $\mathbb{C} \subseteq \mathbb{PS}$.*
2. *If the dominance relation > underlying $\geq_p$ is a direct dominance relation and if the Jackson-Wolinsky rules hold, then $\mathbb{C} = \mathbb{PS}$ and $\mathbb{PS}$ is nonempty if and only if there exists a basin of attraction containing a single network.*

Proof. The proof of part 1 follows from part 1 of Theorem 5 and the fact that $\mathbb{SS}$

[29] In particular, under Jackson-Wolinsky rules, if

$G \rightarrow_S G'$,

then there are only three possibilities:

(i) $G' = G \cup (a, (i, i'))$for some $a \in A$ and $S = \{ i, i'\}$;
(ii) $G \setminus (a, (i, i'))$for some $a \in A$ and $S = \{i\}$; or
(iii) $G \setminus (a, (i, i'))$ for some $a \in A$ and $S = \{i'\}$.

Thus, under Jackson-Wolinsky rules, if a network is not strongly stable, automatically it is not pairwise stable – and thus under Jackson-Wolinsky rules

$\mathbb{PS} \subseteq \mathbb{SS}$.

the path dominance is induced by a direct dominance and if the Jackson-Wolinsky rules hold, then we have $\mathbb{C} = \mathbb{SS} = \mathbb{PS}$. By part 1 of Theorem 4, $\mathbb{C} = \mathbb{SS} = \mathbb{PS}$ is nonempty if and only if there exists a basin of attraction containing a single network. ■

Theorem 6 can be viewed as an extension of a result due Jackson and Watts (2002) on the existence of pairwise stable linking networks for network formation games induced by Jackson-Wolinsky rules. In particular, Jackson and Watts (2002) show that for this particular class of Jackson-Wolinsky network formation games, if there does not exist a closed cycle of networks, then there exists a pairwise stable network. Our notion of a strategic basin of attraction containing *multiple* networks corresponds to their notion of a closed cycle of networks. Thus, stated in our terminology, Jackson and Watts show that for this class of network formation games, if there does not exist a basin of attraction containing multiple networks, then there exists a pairwise stable network. Following our approach, if we specialize to this class of Jackson-Wolinsky network formation games, then by part 2 of Theorem 6 the existence of *at least one* strategic basin containing a single network is both necessary and sufficient for the existence of a pairwise stable network.

5.3 Consistent networks

We begin with a formal definition of farsighted consistency (Chwe, 1994).

Definition 7 – Consistent sets. Let $(\mathbb{G}, \geq_p)$ be a network formation game where path dominance $\geq_p$ is induced by an indirect dominance relation $\triangleright\triangleright$. A subset $\mathbb{F}$ of directed networks in $\mathbb{G}$ is said to be consistent in $(\mathbb{G}, \geq_p)$ if

for all $G_0 \in \mathbb{F}$,
$G_0 \rightarrow_{S_1} G_1$ for some $G_1 \in \mathbb{G}$ and some coalition S_1 implies that
there exists $G_2 \in \mathbb{F}$
with $G_2 = G_1$ or $G_2 \triangleright\triangleright G_1$ such that,
$G_0 \not<_{S_1} G_2$.

In words, a subset of directed networks $\mathbb{F}$ is said to be consistent in $(\mathbb{G}, \geq_p)$ if given any network $G_0 \in \mathbb{F}$ and any deviation to network $G_1 \in \mathbb{G}$ by coalition S_1 (via adding, subtracting, or replacing arcs in accordance with effectiveness relations $\rightarrow_S$), there exists further deviations leading to some network $G_2 \in \mathbb{F}$ where the initially deviating coalition S_1 is not better off – and possibly worse off. A network $G \in \mathbb{G}$ is said to be consistent if $G \in \mathbb{F}$ where $\mathbb{F}$ is a consistent set in $(\mathbb{G}, \geq_p)$.

There can be many consistent sets in $(\mathbb{G}, \geq_p)$. We shall denote by $\mathbb{F}^*$ the *largest consistent set*. Thus, if $\mathbb{F}$ is a consistent set, then $\mathbb{F} \subseteq \mathbb{F}^*$. By Proposition 1 in

Chwe (1994) there exists uniquely a largest consistent set in $(\mathbb{G}, \geq_p)$. Moreover, by the Corollary to Proposition 2 in Chwe (1994) this largest consistent set is nonempty and externally stable with respect to indirect dominance $\rhd\rhd$. This Theorem is essentially a network rendition of the Corollary to Proposition 2 in Chwe (1994).[30]

We now have our main result on the relationship between basins of attraction, stable sets, the path dominance core, and the largest consistent set.

Theorem 7– Basins of attraction, the path dominance core, and the largest consistent set. *Given primitives* $(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S \in P(D)\}}$ *and given network formation game* $(\mathbb{G},\geq_p)$, *where path dominance is induced by an indirect dominance relation* $\rhd\rhd$, *assume without loss of generality that* $(\mathbb{G}, \geq_p)$ *has nonempty largest consistent set given by* $\mathbb{F}^*$ *and basins of attraction given by*

$$\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\}.$$

Then the following statements are true:

1. *Each basin of attraction* $\mathbb{A}_k$, $k = 1,2, \ldots, m$, *has a nonempty intersection with the largest consistent set* $\mathbb{F}^*$, *that is*

   $\mathbb{F}^* \cap \mathbb{A}_k \neq \varnothing$, *for* $k = 1,2, \ldots,m$.

2. *If* $(\mathbb{G}, \geq_p)$ *has a nonempty path dominance core* $\mathbb{C}$, *then*

   $\mathbb{C} \subseteq \mathbb{F}^*$.

Proof. In light of Theorem 4, (2) easily follows from (1). Thus, it suffices to prove (1). Suppose that for some basin of attraction $\mathbb{A}_{k'}$

$$\mathbb{F}^* \cap \mathbb{A}_{k'} = \varnothing.$$

Let G' be a network in $\mathbb{A}_{k'}$. Because $\mathbb{F}^*$ is externally stable with respect to the indirect dominance relation $\rhd\rhd$, $G' \notin \mathbb{F}^*$ implies that there exists some network $G^* \in \mathbb{F}^*$ such that $G^* \rhd\rhd G'$. Thus, $G^* \geq_p G'$. Because the networks in $\mathbb{A}_{k'}$ are without descendants, it must be true that $G' \geq_p G^*$. But this implies that $G^* \equiv_p G'$, and therefore that $G^* \in \mathbb{A}_{k'}$, a contradiction. ■

Remark 3. Recently, Herings et al. (2006) introduced a notion of pairwise farsighted stability. If in our model coalitional preferences $\{\succ_S\}_{S \in P(D)}$ over networks are based on

[30] Page and Kamat (2005) provide an alternative proof of the nonemptiness and external stability of the largest consistent set (with respect to indirect dominance). In particular, Page and Kamat modify the indirect dominance relation so as to make it transitive as well as irreflexive. They then show that the unique stable set with respect to path dominance induced by this new transitive indirect dominance relation is contained in the largest consistent set – and in this way show that the largest consistent set is nonempty and externally stable.

weak preferencerelations $\{\succsim_d\}_{d\in D}$ (see Remark 1 above), if nodes represent players (i.e., $N = D$), and if the dominance relation underlying the path dominance relation is indirect, then under Jackson-Wolinsky rules the corresponding weak path dominance core is contained in the set of pairwise farsightedly stable networks.

5.4 Nash networks

Definition 8 – Nash networks. Given primitives $(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S\in P(D)}$ and network formation game $(\mathbb{G}, \geq_p)$, network $G \in \mathbb{G}$ is said to be a Nash network in $(\mathbb{G}, \geq_p)$ if for all $G' \in \mathbb{G}$ and $S \in P(D)$ such that $|S| = 1$, $G \rightarrow_S G'$ implies that $G \not\prec_S G'$.

Thus, a network is Nash if whenever an individual player has the power to change the network to another network, the player will have no incentive to do so. We shall denote by $\mathbb{NE}$ the set of Nash networks. Note that our definition of a Nash network does not require that the network formation rules, as represented via the effectiveness relations $\{\rightarrow_S\}_{S\in P(D)}$, be noncooperative (see subsection 3.2.1). Also, note that under our definition any network that cannot be changed to another network by a coalition of size 1 is a Nash network. Finally, note that the set of strongly stable networks $\mathbb{SS}$ is contained in the set of Nash networks $\mathbb{NE}$.

We now have our main result on the path dominance core and Nash networks.

Theorem 8 – The path dominance core and Nash networks. *Given primitives $(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S\in P(D)}$ and network formation game $(\mathbb{G}, \geq_p)$, where path dominance $\geq_p$ is induced by either a direct dominance relation or an indirect dominance relation, the following statements are true.*

1. *If the path dominance core $\mathbb{C}$ of $(\mathbb{G}, \geq_p)$ is nonempty, then $\mathbb{NE}$ is nonempty and $\mathbb{C} \subseteq \mathbb{NE}$.*
2. *If the dominance relation $>$ underlying $\geq_p$ is a direct dominance relation and if the rules of network formation are such that $G \rightarrow_S G'$ implies that $|S| = 1$, then $\mathbb{C} = \mathbb{NE}$ and $\mathbb{NE}$ is nonempty if and only if there exists a basin of attraction containing a single network.*

Proof. The proof of part 1 follows from part 1 of Theorem 5 and the fact that $\mathbb{SS} \subseteq \mathbb{NE}$. For the proof of part 2, note that if the rules of network formation are such that $G \rightarrow_S G'$ implies that $|S| = 1$, then $\mathbb{SS} = \mathbb{NE}$. Thus, we have $\mathbb{C} \subseteq \mathbb{SS} = \mathbb{NE}$. If in addition the path dominance relation is induced by a direct dominance relation, then we have $\mathbb{NE} = \mathbb{SS} \subseteq \mathbb{C}$, and we conclude that $\mathbb{C} = \mathbb{SS} = \mathbb{NE}$. Thus, if the path dominance is induced by a direct dominance and if the rules are such that $G \rightarrow_S G'$ implies that $|S| = 1$, then we have $\mathbb{C} = \mathbb{SS} = \mathbb{NE}$. By part 1 of Theorem 4, $\mathbb{C} = \mathbb{SS} = \mathbb{NE}$ is nonempty if and only if there exists a basin of attraction containing a single network. ■

We close this section by noting that if the dominance relation > underlying $\geq_p$ is a direct dominance relation and if the rules of network formation are such that $G \rightarrow_S G'$ implies that $|S| = 1$, then the set of Nash networks $\mathbb{NE}$ is contained in the set of constrained Pareto efficient networks $\mathbb{E}$. Thus, for this case we have $\mathbb{C} = \mathbb{SS} = \mathbb{NE} \subseteq \mathbb{E}$.

6. Examples

In the abstract games, $(\mathbb{G}, \geq_p)$, that we have considered, the set of outcomes $\mathbb{G}$ is a set of directed networks and we have focused on path dominance induced by either direct dominance or indirect dominance. However, our main results, Theorems 1-4, hold for any abstract game with a finite set of outcomes equipped with path dominance induced by any dominance relation. With this in mind, in this section we will demonstrate the flexibility of our approach and the wide applicability of our results by first considering network games with a potential function and then considering games where the set of outcomes is, in one case, a set of linking networks and, in another case, a set of coalition structures, and where the path dominance relation is induced by a dominance relation other than a direct or an indirect dominance relation (as defined in sections 3.3.1 and 3.3.2). In particular, in our first example, we consider noncooperative network formation games and show that any noncooperative network formation game possessing a potential function has basins of attraction each consisting of a single network and thus has a nonempty path dominance core. In our second example, we show how our approach can be applied to Jackson-Wolinsky linking networks and we provide necessary and sufficient conditions for nonemptiness of the set of pairwise stable linking networks. Finally, we show via an example proposed to us by Salvador Barbera and Michael Maschler (2006) how our approach can be used to analyze hedonic games, and in particular, we show how farsightedness can lead to instability (i.e., emptiness of the path dominance core) in hedonic games.

6.1 Noncooperative network formation games possessing a potential function

Suppose the primitives $(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S \in P(D)}$ underlying the network formation game $(\mathbb{G}, \geq_p)$ are such that:

1. the set of nodes N and the set of players D are one and the same (i.e., $N = D$ and $\mathbb{G} \subseteq P(A \times (D \times D))$);
2. preferences $\{\succ_S\}_{S \in P(D)}$ over networks $\mathbb{G}$ are specified via player payoff functions $v_d(\cdot)$, that is, coalition $S' \in P(D)$ prefers network G' to network G if $v_d(G') > v_d(G)$ for all $d \in S'$;[31]

[31] This is a frequently used way of defining payoffs to coalitions; see for example, Jackson (2005) and van den Nouweland (2005).

3. effectiveness relations $\{\rightarrow_S\}_{S\in P(D)}$ over networks $\mathbb{G}$ are such that,
   (i) adding an arc a from player i to player i' requires only that player i agree to add the arc (i.e., arc addition is unilateral and can be carried out only by the initiator, player i),
   (ii) subtracting an arc a from player i to player i' requires only that player i agree to subtract the arc (i.e., arc subtraction is unilateral and can be carried out only by the initiator, player i), and
   (iii) $G \rightarrow_S G'$ implies that $|S| = 1$ (i.e., only network changes brought about by individual players are allowed);
4. the dominance relation > over $\mathbb{G}$ is given by a direct dominance relation $\triangleright$, that is, $G' \triangleright G$ if and only if for some player $d' \in D$, $v_{d'}(G') > v_{d'}(G)$ and $G \rightarrow_{d'} G'$.

We say that the noncooperative network formation game $(\mathbb{G}, \geq_p)$ is a potential game if there exists a function

$$P(\cdot)\colon \mathbb{G} \rightarrow R$$

such that for all G and G' with $G \rightarrow_{d'} G'$ for some player d',

$$v_{d'}(G') > v_{d'}(G) \text{ if and only if } P(G') > P(G).$$

It is easy to see that any noncooperative network formation game $(\mathbb{G}, \geq_p)$ possessing a potential function has no circuits, and thus possesses strategic basins of attraction each consisting of a single network without descendants.[32] Thus, we can conclude from our Theorem 4 that any noncooperative network formation game possessing a potential function has a nonempty path dominance core. In addition, we know from our Theorem 8 that in this example the path dominance core $\mathbb{C}$ is equal to the set of Nash networks $\mathbb{NE}$.[33]

6.2 Jackson-Wolinsky linking networks

Consider primitives $(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S\in P(D)}$ with corresponding network formation game $(\mathbb{G}, \geq_p)$ where $\mathbb{G}$ is given by a feasible set of linking networks, coalitional preferences $\{\succ_S\}_{S\in P(D)}$ are based on weak preferences (see Remarks 1 and 2 above), effectiveness relations $\{\rightarrow_S\}_{S\in P(D)}$ are specified via Jackson-Wolinsky rules, and the dominance relation > is direct. In particular, assume that the set of nodes N and

[32] As has been shown by Monderer and Shapley (1996), potential games are closely related to congestion games introduced by Rosenthal (1973) – also see Holzman and Law-Yone (1997).

[33] Page and Wooders (2007b) introduce a club network formation game which is a variant of the noncooperative network formation game described above and show that this game possesses a potential function (see also Page and Wooders 2007a). Prior papers studying potential games in the context of linking networks include Slikker et al. (2000) and Slikker and van den Nouweland (2002). These papers have focused on providing the strategic underpinnings of the Myerson value (Myerson 1977).

the set of players D are equal, let g^N denote the collection of all subsets of N of size 2, and let $\mathbb{G}$ be a nonempty subset of $P(g^N)$, where $P(g^N)$ denotes the collection of all nonempty subsets of g^N (i.e., the set of all linking networks – see the definition in Jackson and Wolinsky, 1996). To simplify comparisons, we use the standard notation for linking networks and let g denote a typical linking network.

Under Jackson-Wolinsky rules, if $g \rightarrow_{S'} g'$ then $g \neq g'$ and either (i) $g' = g \cup \{i, i'\}$ (a link is added between players i and i') and $S' = \{i, i'\}$ or (ii) $g' = g \backslash \{i, i'\}$ (the link between players i and i' is removed) and $S' = \{i\}$ or $S' = \{i'\}$ or $S' = \{i, i'\}$. Moreover, if coalitional preferences $\{\succ_S\}_{S \in P(D)}$ are based on weak preference relations $\{\succsim_d\}_{d \in D}$, then coalition $S' \in P(D)$ *prefers* network g' to network g, written $g' \succ_{S'} g$, if for all players $d \in S'$, $g' \succsim_d g$ and if for at least one player $d' \in S'$, $g' \succ_{d'} g$.[34] Finally, if the dominance relation > is direct with underlying weak preferences, then $g' \rhd g$ if and only if either (i) $g \rightarrow_{\{i, i'\}} g'$ and $g' \succ_{\{i, i'\}} g$ where $g' = g \cup \{i, i'\}$or (ii) (a) $g \rightarrow_{\{i\}} g'$ and $g' \succ_{\{i\}} g$ where $g' = g \backslash \{i, i'\}$ or (b) $g \rightarrow_{\{i'\}} g'$ and $g' \succ_{\{i'\}} g$ where $g' = g \backslash \{i, i'\}$.

It follows from our Theorem 4 that any network formation game $(\mathbb{G}, \geq_p)$ induced by Jackson-Wolinsky primitives (i.e., primitives as specified above) has a nonempty path dominance core if and only if there is at least one strategic basin of attraction containing a single network. Moreover, it follows from our Theorem 6 that for any such network formation game the path dominance core is equal to the set of pairwise stable networks (as defined for linking networks in Jackson and Wolinsky, 1996).

6.3 Hedonic games

Consider a hedonic game where a move from one coalition structure to another can be initiated by any group of players defecting from the original structure, but in order for the change to prevail all players in coalitions augmented or created by the defecting players must prefer their new coalitions to their old coalitions – or must prefer their eventual coalitions to their old coalitions if players are farsighted. Call the path dominance core with respect to direct dominance, the hedonic direct core and the path dominance core with respect to indirect dominance the hedonic farsighted core. Note that the hedonic direct core is equivalent to the usual hedonic core. As the following example will show, the hedonic farsighted core may be empty even when the hedonic core is not.

Consider the following hedonic game proposed to us by Barbera and Maschler (2006). Let the player set be given by $D = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Player preferences over coalitions are as follows:

[34] Recall from Remark 1 that if $g' \succsim_d g$ then player d either strictly prefers g' to g (denoted $g' \succ_d g$) or is indifferent between g' and g (denoted $g' \sim_d g$).

Table 1. Players' preferences over coalitions

player 1	(1, 2, 3, 4)	(1, 2, 3)	(1, 2)	(1)	…
player 2	(1, 2, 3, 4)	(1, 2, 3)	(1, 2)	(2)	…
player 3	(1, 2, 3, 4)	(3, 4, 5, 6)	(1, 2, 3)	(3)	(3, 6)
player 4	(1, 2, 3, 4)	(3, 4, 5, 6)	(4, 5)	(4)	…
player 5	(3, 4, 5, 6)	(5, 6, 7, 8)	(4, 5)	(5)	…
player 6	(3, 4, 5, 6)	(5, 6, 7, 8)	(6, 7, 8)	(6)	(3, 6)
player 7	(5, 6, 7, 8)	(6, 7, 8)	(7, 8)	(7)	…
player 8	(5, 6, 7, 8)	(6, 7, 8)	(7, 8)	(8)	…

Consider the row for player 1 in the table above. The interpretation is that 1 prefers the coalition (1, 2, 3, 4) to the coalition (1, 2, 3), to the coalition (1, 2), and so on. Player 1's preferences over the remaining coalitions are irrelevant to the following example so they are not specified. The same interpretation applies to the rows corresponding to other players.

A partition of the player set is in the *hedonic core* if there does not exist a coalition that is preferred by all its members to their coalitions of membership in the original partition (i.e., a partition is in the hedonic core if it is not directly dominated by another partition). Consider the partition ((1, 2, 3, 4), (5, 6, 7, 8)) $\in \mathbb{G}$. This is a core point for the hedonic game because the only coalition that is preferred by players 5 and 6 is (3, 4, 5, 6) but two members of this coalition, 3 and 4, do not prefer it (i.e., ((1, 2), (3, 4, 5, 6), (7, 8)) does not directly dominate ((1, 2, 3, 4), (5, 6, 7, 8))). If players 4 and 5 are farsighted, however, and domination is indirect, 4 and 5 can decide to form a coalition (4, 5) – thus bringing about the partition ((1, 2, 3), (4, 5), (6, 7, 8)). Now players 3,4,5, and 6 could all benefit from forming a coalition. This brings us to the partition ((1, 2), (3, 4, 5, 6), (7, 8)) a hedonic core point in which 4 and 5 are better off than in the original hedonic core point. Thus, ((1, 2), (3, 4, 5, 6), (7, 8)) indirectly dominates ((1, 2, 3, 4), (5, 6, 7, 8)). But the story is not finished. Starting from ((1, 2), (3, 4, 5, 6), (7, 8)), players 3 and 6 can separate and form their own coalition. Using an argument similar to the one above, this move by 3 and 6 can then lead back to the original partition. Thus, ((1, 2, 3, 4), (5, 6, 7, 8)) indirectly dominates ((1, 2), (3, 4, 5, 6), (7, 8)).

We see here that, even though the hedonic core is nonempty, the hedonic farsighted core is empty. Another point illustrated is that for path dominance, it is only necessary that a coalition perceive *some* path that would lead to a preferred situation; it is not required that a coalition perceive some preferred *final* (and presumably stable) outcome. The example also suggests for those special cases of hedonic games where the hedonic direct core (i.e., the hedonic core) is non-empty and not a singleton, then the path dominance core with respect to indirect dominance (i.e., the hedonic farsighted core) is empty. (See Diamantoudi and Xue, 2003 for related work applying indirect dominance to hedonic games).[35]

7. Conclusions

From the viewpoint of the path dominance core with direct or indirect dominance, there are a number of potential questions to be addressed. For example, what is the relationship, if any, between basins of attraction and the path dominance core and partnered (or separating) collections of coalitions, as in for example Page and Wooders (1996), Reny and Wooders (1997) or Maschler and Peleg (1967) and Maschler et al. (1971)? Or what is relationship between basins of attraction and the path dominance core and the inner core, as in Qin (1993,1994)?

To conclude, we return to the prior research introducing concepts similar to the abstract game defined in this paper and the union of basins of attractions; see Schwartz (1974), Kalai et al. (1976), Kalai and Schmeidler (1977) and Shenoy (1980).[36] For specificity, we focus on Kalai and Schmeidler (1977). These authors take as given a set of feasible alternatives, denoted by S, a dominance relation, denoted by M and the transitive closure of M, denoted by $\hat{M}$. Their admissible set is the set $A(S, M) := \{x \in S\colon y \in S$ and $y\hat{M}x$ imply $x\hat{M}y\}$.[37]

Besides non-emptiness of the admissible set, they also shown that the admissible set is equal to the union of certain subsets – in our terminology, basins of attraction. While Kalai and Schmeidler apply their concept to cooperative games and games in normal (strategic) form, they do not consider networks, the focus of our research. Once our model of network formation is developed, then our abstract game is a particular case of the abstract game of these earlier authors. Our contribution differs in that we develop the network framework and characterize several equilibrium concepts from network theory in terms of their relationships to each other and to basins of attraction and the path dominance core. In addition, we characterize the set of von-Neumann-Morgenstern solutions and the path-dominance core (a case of the abstract core notion introduced in Gilles 1959) in terms of their relationships to basins of attraction. It may well be that the insightful examples developed by these authors will lead to new sorts of

[35] In brief, the effectiveness relations in Diamantoudi and Xue differ from the effectiveness relations in our rendition of the Barbera-Maschler example. In particular, in Diamantoudi and Xue all defecting players must form a coalition in the new partition, whereas in the Barbera-Maschler example, defecting players can join already existing coalitions in forming the new partition. Moreover, in Diamantoudi and Xue only defecting players must prefer their new coalition in order for the change to take place, whereas in the Barbera-Maschler example, not only must defecting players prefer their new coalitions, but also all players in coalitions joined by the defecting players must prefer their new coalitions in order for the change to take place.

[36] We thank Sylvie Thoron for bringing this to our attention.

[37] Kalai and Schmeidler (1977) also cite Schwartz (1974) for the origins of this concept.

examples for networks, a question we are currently addressing. Also, Kalai and Schmeidler (1977) allow an infinite set of possibilities, which, in a network framework, introduces a host of new questions. We plan to address some of these in future research.

References

Aumann, R.J. (1964), 'Markets with a continuum of traders', *Econometrica* **32**, 39–50.

Bala, V. and S. Goyal (2000), 'A noncooperative model of network formation', *Econometrica* **68**, 1181–1229.

Barbera, S. and M. Maschler (2006), 'Private correspondence'.

Berge, C. (2001), *The Theory of Graphs*, Mineola, NY: Dover (reprint of the translated French edition published by Dunod, Paris, 1958).

Calvó-Armengol, A. and R. Ilkilic (2005), 'Pairwise stability and Nash equilibria in network formation', FEEM Working Paper 34.05.

Chwe, M. (1994), 'Farsighted coalitional stability', *Journal of Economic Theory* **63**, 299–325.

Demange, G. (2004), 'On group stability and hierarchies in networks', *Journal of Political Economy* **112**, 754–778.

Diamantoudi, E. and L. Xue (2003), 'Farsighted stability in hedonic games', *Social Choice and Welfare* **21**, 39–61.

Gillies, D.B. (1959), 'Solutions to general non-zero-sum games', in A.W. Tucker and R.D. Luce (eds.), *Contributions to the Theory of Games*, Volume 4, Princeton University Press.

Guilbaud, G.T. (1949), 'La theorie des jeux', *Économie appliquée* **2**, p. 18.

Harsanyi, J.C. (1974), 'An equilibrium-point interpretation of stable sets and a proposed alternative definition', *Management Science* **20**, 1472–1495.

Herings, P. J.-J., A. Mauleon and V. Vannetelbosch (2006), 'Farsightedly stable networks', Meteor Research Memorandum RM/06/041.

Holzman, R. and N. Law-Yone (1997), 'Strong equilibrium in congestion games', *Games and Economic Behavior* **21**, 85–101.

Inarra, E., J. Kuipers and N. Olaizola (2005), 'Absorbing and generalized stable sets', *Social Choice and Welfare* **24**, 433–437.

Jackson, M.O. (2005), 'A survey of models of network formation: Stability and efficiency', in G. Demange and M.H. Wooders (eds.), *Group Formation in Economics: Networks, Clubs, and Coalitions*, Cambridge: Cambridge University Press.

Jackson, M.O. and A. van den Nouweland (2005), 'Strongly stable networks', *Games and Economic Behavior* **51**, 420–444.

Jackson, M.O. and A. Watts (2002), 'The evolution of social and economic networks', *Journal of Economic Theory* **106**, 265–295.

Jackson, M.O. and A. Wolinsky (1996), 'A strategic model of social and economic networks', *Journal of Economic Theory* **71**, 44–74.

Kalai, E., A. Pazner and D. Schmeidler (1976), 'Collective choice correspondences as admissible outcomes of social bargaining processes', *Econometrica* **44**, 233–240.

Kalai, E. and D. Schmeidler (1977), 'An admissible set occurring in various bargaining situations', *Journal of Economic Theory* **14**, 402–411.

Lucas, W.F. (1968), 'A game with no solution', *Bulletin of the American Mathematical Society* **74**, 237–239.

Maschler, M. and B. Peleg (1967), 'The structure of the kernel of a cooperative game, *SIAM Journal of Applied Mathematics* **15**, 569–604.

Maschler, M., B. Peleg and L.S. Shapley (1971), 'The kernel and bargaining set for convex games', *International Journal of Game Theory* **1**, 73–93.

Mauleon, A. and V. Vannetelbosch (2003), 'Farsightedness and cautiousness in coalition formation', FEEM Working Paper 52.03.

Monderer, D. and L.S. Shapley (1996), 'Potential games', *Games and Economic Behavior* **14**, 124–143.

Moulin, H. and B. Peleg (1982), 'Cores of effectivity functions and implementation theory', *Journal of Mathematical Economics* **10**, 115–145.

Myerson, R.B. (1977), 'Graphs and cooperation in games', *Mathematics of Operations Research* **2**, 225–229.

Page Jr., F.H. and S. Kamat (2005), 'Farsighted stability in network formation', in G. Demange and M.H. Wooders (eds.), *Group Formation in Economics: Networks, Clubs, and Coalitions*, Cambridge: Cambridge University Press.

Page Jr., F.H. and M.H. Wooders (1996), 'The partnered core and the partnered competitive equilibrium', *Economics Letters* **52**, 143–152.

Page Jr., F.H. and M.H. Wooders (2007a), 'Networks and clubs', *Journal of Economic Behavior and Organization* **64**, 406–425.

Page Jr., F.H. and M.H. Wooders (2007b), 'Club networks with multiple memberships and noncooperative stability', Indiana University typescript, paper presented at the International Conference on the Formation of Social Networks, Paris, Carre des Sciences, June 28–29, 2007.

Page Jr., F.H., M.H. Wooders and S. Kamat (2005), 'Networks and farsighted stability', *Journal of Economic Theory* **120**, 257–269.

Qin, C.-Z. (1993), 'A conjecture of Shapley and Shubik on competitive outcomes in the cores of NTU market games', *International Journal of Game Theory* **22** , 335–344.

Qin, C.-Z. (1994), 'The inner core of an N-person game', *Games and Economic Behavior* **6**, 431–444.

Reny, P.J. and M.H. Wooders (1996), 'The partnered core of a game without side payments', *Journal of Economic Theory* **70**, 298–311.

Richardson, M. (1953), 'Solutions of irreflexive relations', *Annals of Mathematics* **58**, 573–590.

Rosenthal, R.W. (1972), 'Cooperative games in effectiveness form', *Journal of Economic Theory* **5**, 88–101.

Rosenthal, R.W. (1973), 'A class of games possessing pure-strategy Nash equilibria', *International Journal of Game Theory* **2**, 65–67.

Scarf, H. (1967), 'The core of an N-person game', *Econometrica* **35**, 50–69.

Schwartz, T. (1974), 'Notes on the abstract theory of collective choice', School of Urban and Public Affairs, Carnegie-Mellon University.

Shenoy, P.P. (1980), 'A dynamic solution concept for abstract games', *Journal of Optimization Theory and Applications* **32**, 151–169.

Slikker, M., B. Dutta, A. van den Nouweland and S. Tijs (2000), 'Potential maximizers and network formation', *Mathematical Social Sciences* **39**, 55–70.

Slikker, M. and A. van den Nouweland (2002), 'Network formation, costs, and potential games', in P. Borm and H. Peters (eds.), *Chapters in Game Theory*, Kluwer Academic Publishers.

van den Nouweland, A. (2005), 'Models of network formation in cooperative games', in G.

Demange and M.H. Wooders (eds.), *Group Formation in Economics: Networks, Clubs, and Coalitions*, Cambridge: Cambridge University Press.

von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Xue, L. (1998), 'Coalitional stability under perfect foresight', *Economic Theory* **11**, 603–627.

Network Games

Andrea Galeotti, Sanjeev Goyal, Matthew O. Jackson, Fernando Vega-Redondo and Leeat Yariv

Paper presented at the CTN Workshop 2006. Reprinted from *Review of Economic Studies* **77**(1), January 2010, 218–244, with permission from Oxford Journals.

In contexts ranging from public goods provision to information collection, a player's well-being depends on his or her own action as well as on the actions taken by his or her neighbours. We provide a framework to analyse such strategic interactions when neighbourhood structure, modelled in terms of an underlying network of connections, affects payoffs. In our framework, individuals are partially informed about the structure of the social network. The introduction of incomplete information allows us to provide general results characterizing how the network structure, an individual's position within the network, the nature of games (strategic substitutes vs. complements and positive vs. negative externalities) and the level of information shape individual behaviour and payoffs.

1. Introduction

In a range of social and economic interactions – including public goods provision, job search, political alliances, trade, friendships and information collection – an agent's well-being depends on his or her own actions as well as on the actions taken by his or her neighbours. For example, the decision of an agent of whether or not to buy a new product, or to attend a meeting, is often influenced by the choices of his or her friends and acquaintances (be they social or professional).

We thank the editor and three anonymous referees for useful suggestions. We are also grateful to Willemien Kets and a number of seminar audiences for comments which significantly improved the quality and broadened the scope of the paper. Jackson gratefully acknowledges financial support from the Center for Advanced Studies in the Behavior Sciences, the Guggenheim Foundation and the NSF under grant SES-0647867. Vega-Redondo gratefully acknowledges financial support from the Spanish Ministry of Education under grant SEJ2007-62656.

The empirical literature identifying the effects of agents' neighbourhood patterns (i.e. their social network) on behaviour and outcomes has grown over the past several decades.[1] The emerging empirical evidence motivates the theoretical study of network effects. We would like to understand how the pattern of social connections shapes the choices that individuals make and the payoffs they can hope to earn. We would also like to understand how changes in the network matter as this tells us how individuals would like to shapes the networks in which they are located.

Attempts at the study of these basic questions have been thwarted by a fundamental theoretical problem: even the simplest games played on networks have multiple equilibria, which display a bewildering range of possible outcomes. The literature on global games illustrates how the introduction of (a small amount of) incomplete information can sometimes resolve the problem of multiplicity as well as provide interesting and novel economic intuitions.[2] Recently, this approach has faced the critique that the equilibrium selection achieved depends on the specifics of the incomplete information that is assumed, a point made convincingly by Weinstein and Yildiz (2007). However, in the context of network games there is a natural way to introduce incomplete information that eliminates this ambiguity, which is having uncertainty about the identity of players' future neighbours and the number of neighbours that they will have. There are many decisions that are made at times where a player has a good forecast of the number of her connections (her degree) but has incomplete information about the degrees of others.[3]

Indeed, in many circumstances individuals are aware of their proclivity to interact with others, but do not know who these partners are at the time of choosing actions. For instance, students who are planning a career of international diplomacy may anticipate how many individuals each of them will most likely interact with, but do not know who these individuals will be when deciding the number of foreign languages to study; or researchers choosing software based on compatibility may know the number of coauthors they expect to have in the future, but not necessarily who these people will be; or individuals deciding whether to get a medical vaccine may anticipate the volume of people they will interact with, but not specifically who these people will be. For these kind of environments, our model highlights the following two features: (i) agents have a good sense of the volume of agents each of them will interact with (their respective degree); and (ii) action choices are taken prior to the actual network of

[1] The literature is much too vast to survey here; influential works include Katz and Lazarsfeld (1955), Coleman (1966), Granovetter (1994), Foster and Rosenzweig (1995), Glaeser et al. (1996), Topa (2001) and Conley and Udry (2010).

[2] Starting with the work of Carlsson and van Damme (1993), there is now an extensive literature on global games. For a survey of this work, see Morris and Shin (2003).

[3] For discussion of the knowledge of individuals about the network see, e.g. Kumbasar et al. (1994).

connections being realized (i.e. there is incomplete information regarding the identity of neighbours, neighbours' neighbours, etc.).

Motivated by these considerations we develop a model of games played on networks, in which players have private and incomplete information about the network. Their private knowledge about the network is interpreted as their type, and we study the Bayes – Nash equilibria of this game. We find that much of the equilibrium multiplicity that arises under complete information is no longer sustainable under incomplete information. Specifically, the key insight is that when players have limited information about the network they are unable to condition their behaviour on its fine details and this leads to a significant simplification and sharpening of equilibrium predictions.

There are two other important aspects of our framework that we would like to stress here. One is that individuals are allowed to have beliefs about degrees of their neighbours that depend on their own degree. We capture correlations in the degrees of neighbours through a weakening of the notion of affiliation, which is a measure widely used in economics to capture joint correlations in types. As we explain below, many of the real world contexts studied by the network literature display degree correlations (positive or negative) that fall into one of the scenarios considered here. This is also true for much of the theoretical work concerned with alternative models of network formation.

A second important feature of our approach is that we allow for alternative scenarios on how a player's payoffs are affected by the actions of others. This is motivated by our desire to develop an understanding of how the payoffs interact with the network structure. We focus, therefore, on two canonical types of interaction: strategic complements and strategic substitutes.[4] These two cases cover many of the game-theoretic applications studied by the economic literature.

We now provide an overview of our main results. Our first result shows the existence of an equilibrium involving monotone (symmetric) strategies. In particular, in the case of strategic substitutes equilibrium actions are non-increasing in players' degrees, whereas under strategic complements equilibrium actions are non-decreasing in players' degrees. We also provide conditions under which all (symmetric) equilibria are monotone. In turn, the monotonicity property of equilibrium actions implies that social connections create personal advantages irrespective of whether the game exhibits strategic complements or substitutes: in games with positive externalities well-connected players earn more than poorly

[4] For instance, strategic complements arise whenever the benefit that an individual obtains from buying a product or undertaking a given behaviour is greater as more of his partners do the same. This might be due to direct effects of having similar or compatible products (such as in the case of computer operating systems), peer pressures (as in the case of drug use) and so forth. The strategic substitutes case encompasses many scenarios that allow for free riding or have a public good structure of play, such as costly experimentation or information collection. Formal definitions of these games are given in Section 3.

connected players.[5] This provides a first illustration of the additional structure afforded by our assumption of incomplete information. Building upon it, our second objective is to understand how changes in the perceived social network affect equilibrium behaviour and welfare within the different payoff scenarios. We start by considering the effects induced by increased connectivity, as embodied by shifts in the degree distribution that suitably extend the standard notion of first order stochastic dominance (FOSD). This is proven to have unambiguous effects on equilibrium behaviour under strategic substitutes for binary-action games as well as for general games with strategic complements. For binary-action games, we also derive results that involve arbitrary changes in the degree distribution relative to the equilibrium actions. Finally, we explore the implications of endowing agents with deeper (but still local) information on the network. We find that this may lead to non-monotonic equilibrium behaviour.

Our paper is a contribution to the growing literature that, in recent years, has undertaken the study of games played on networks (for an extensive overview of the networks literature, see Goyal (2007) and Jackson (2008)). For instance, decisions to undertake criminal activity (Ballester et al., 2006), public good provision (Bramoullé and Kranton, 2007), the purchase of a product (Galeotti, 2008) and research collaboration among firms (Goyal and Moraga-Gonzalez, 2001) have been studied for specific network structures under complete information.6[6] We would also like to mention Jackson and Yariv (2005), Galeotti and Vega-Redondo (2006) and Sundararajan (2006), who study games with incomplete network knowledge in specific contexts. The principal contribution of our paper is the development of a general framework for the study of games in such an incomplete information setup. We accommodate a large class of games with strategic complements and strategic substitutes, including practically all the applications mentioned above as special cases. Our approach also allows naturally for general patterns of correlations across the degrees of neighbours, and this is important as empirical work suggests that real world networks display such features. To the best of our knowledge, our paper is the first attempt to incorporate general patterns of degree correlations in the study of network games.[7]

There is also a literature in computer science that examines games played on a network; see, e.g., the model of 'graphical games' as introduced by Kearns et al.

[5] The idea that social connections create personal advantages is a fundamental premise of the influential work of Granovetter (1994) and is central to the notion of structural holes developed by Burt (1994). A number of recent empirical studies document the role of connections in providing personal advantages – ranging from finding jobs, getting promotions and gaining competitive advantages in markets.

[6] In particular, regular networks (in which all players have the same degree) and core –periphery structures (the star network being a special case) have been extensively used in the literature.

[7] Jackson and Yariv (2007) follow up on the approach introduced in this paper and obtain complementary results. They examine the multiplicity of equilibria of games on networks with incomplete information, but with a binary action model and a different formulation of payoffs. See also Jackson and Yariv (2008) for a review of related results.

(2001), also analysed by Kakade et al. (2003), among others.[8] The graphical-games literature has focused on the complexity of, and algorithms for, computing equilibria in two-action complete information games played on networks. Here, we allow for more general games and examine different information structures. Importantly, our focus is on the structure of equilibria and its interaction with the underlying network, rather than with the computational complexity of equilibria.

The rest of the paper is organized as follows. In Section 2, we discuss some simple examples that convey many of the insights to be gathered from the general analysis. Our theoretical framework of games played on networks is then introduced in Section 3. Section 4 presents results on the existence and monotonicity of equilibria. Section 5 takes up the study of the effects of network changes on equilibrium behaviour and payoffs. While the analysis in Sections 4 and 5 focuses on a setting in which players know their own degree and have some beliefs about the rest of the network, Section 6 takes up the issues that arise when players have deeper knowledge about the network. Section 7 concludes. All the proofs are gathered in the Appendix.

2. Effects of Networks on Behaviour and Payoffs: Examples

This section presents and analyses two simple games played on networks – reflecting strategic substitutes and strategic complements, respectively – to illustrate the main insights of the paper.

We start with the setting studied by Bramoullé and Kranton (2007) – henceforth referred to as BK. It is a model of the local provision of information (or a local public good) and agents' actions are strategic substitutes. We compare the equilibrium predictions under the assumption of complete information and incomplete information.

Consider a society of n agents, each of them identified with a node in a social network. The links between agents reflect social interactions, and connected agents are said to be 'neighbours'. It is posited that every individual must choose independently an action in $X = \{0, 1\}$, where action 1 may be interpreted as acquiring information, getting vaccinated, etc. and action 0 as not doing so. To define the payoffs, let $y_i \equiv x_i + \overline{x}_{N_i}$ where x_i is the action chosen by agent i, N_i is the set of i's neighbours and $\overline{x}_{N_i} \equiv \sum_{j \in N_i} x_j$ is the aggregate action in N_i. The *gross* payoff to agent i is assumed equal to 1 if $y_i \geq 1$, and 0 otherwise. On the other hand, there is a cost c, where $0 < c < 1$ for choosing action 1, while action 0 bears no cost. Gross payoffs minus costs define the (net) payoffs of the game.[9] Therefore,

[8] There are also models of equilibria in social interactions where players care about the play of certain other groups of players. See Glaeser and Scheinkman (2003) for an overview.

[9] The game is sometimes referred to as the best shot game. For a more detailed presentation of it, see Section 3.

an agent would prefer that someone in his or her neighbourhood take would, however, be willing to take the action 1 if nobody in the neighbourhood does.

We start with the informational assumption made by BK: agents have complete information on the social network and thus the natural equilibrium concept is a Nash equilibrium. It is immediately observed that, as $c < 1$, $y_i \geq 1$ for every player $i \in N$ in any Nash equilibrium. Let us first examine the relation between network connections and actions. In general, such a complete information context allows for a very rich set of Nash equilibria of the induced games. To see this, consider the simple case of a star network and note that there exist two equilibria.

In one equilibrium, the centre chooses 1 and the peripheral players choose 0, whereas in the second equilibrium the peripheral players choose 1 and the centre chooses 0. In the former equilibrium, the centre earns less than the peripheral players, whereas in the latter equilibrium it is the opposite. Figure 1 depicts these possibilities. Hence, even in the simplest networks there exist multiple equilibria and, most importantly, the relation between network connections, equilibrium actions and payoffs may exhibit very different patterns *even when all agents of the same degree choose the same actions*.

Still remaining under the assumption of complete information, note that the effects of adding links to a network on equilibrium actions and aggregate payoffs depend very much on the details of the network and where the links are added (a point made by BK). To see this, consider a network with two stars, each of which contain five peripheral players. Fix a symmetric equilibrium in which the two centres choose action 1, while the peripheral players all choose 0. The aggregate payoff in this equilibrium is $12 - 2c$. In the new network, the old action profile still constitutes an equilibrium. Consider adding a link between the centres of the two stars. In this case the old profile of actions is no longer in equilibrium. In fact, there is *no* equilibrium where both of the original centres choose 1. There is, however, an equilibrium in which the peripheral players of the stars choose 1 and

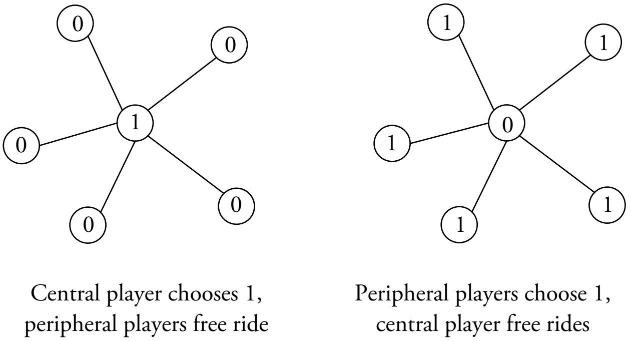

0
0
1
0
0
0
1
1
0
1
1
1
Central player chooses 1,
peripheral players free ride
Peripheral players choose 1,
central player free rides

Figure 1. Strategic substitutes with complete information

the centres choose 0. In this equilibrium there is a clear change in profile of actions, and the aggregate payoffs are given by $12 - 10c$. There is another equilibrium where one of the two centres takes the action 1, and the other does not, and this leads to aggregate payoffs of $12 - 6c$. It follows that in any of the equilibria associated with the addition of the link, aggregate payoffs are lower than in the starting equilibrium. Figure 2 illustrates these outcomes. Interestingly, if a link is added between the centre node of one star and a peripheral player on the other star, as in the bottom of Figure 2, the original equilibrium actions remain part of an equilibrium.

Now let us relax the assumption of complete information on the social network and assume, instead, that players do not know the whole network but are informed only of their own degree. For example, agents' learning may occur prior to the network being realized (say, taking agricultural classes in college prior to opening a winery) or agents may decide to get an immunization (for the flu, hepatitis, etc.) before knowing the individuals they will interact with over the course of the year.

Moreover, assume that players' beliefs about the rest of the network are summarized by a probability distribution over the degrees of their neighbours. For expositional simplicity, suppose also that these beliefs are independent across neighbours as well as of own degree. Under these conditions, a player's (pure) strategy can be identified with a mapping σ specifying the action $\sigma(k) \in X$ chosen for each player of degree k. This game can be studied within the framework of

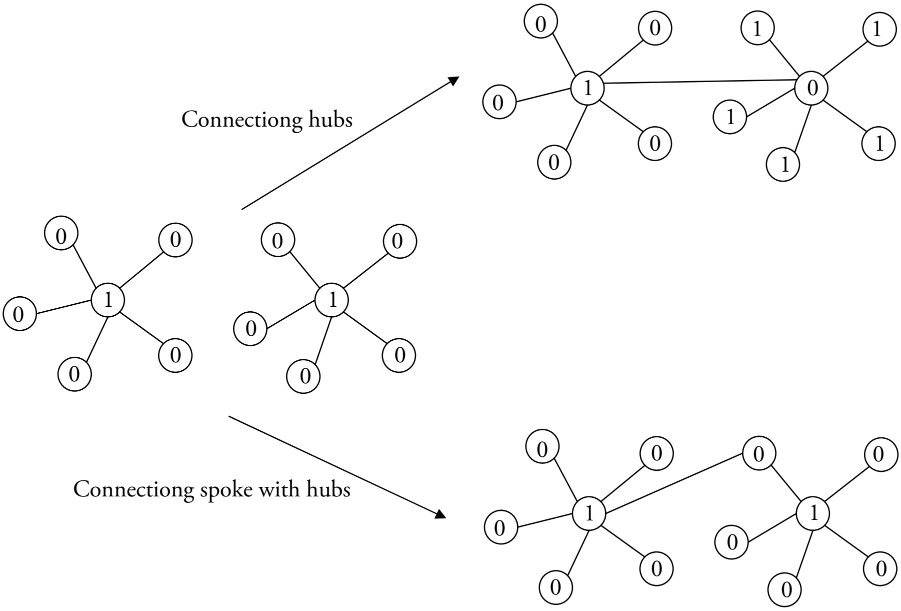


Figure 2. The effects of adding links

Bayesian games of incomplete information by identifying player types with their corresponding degrees.

For concreteness, suppose that a link between any two of n agents is formed independently with probability $p \in (0, 1)$ (commonly referred to as an Erdös–Rényi network (Erdös and Rényi, 1960)). Asymptotically, beliefs about neighbours' degrees then follow a binomial distribution. The probability that any randomly selected neighbour is of degree k is the probability the neighbour is connected to $k - 1$ additional agents of the remaining $n - 2$ agents, and is therefore given by:

$$Q(k;p) = \binom{n-2}{k-1} p^{k-1}(1-p)^{\mathrm{n}-k-1}. \tag{1}$$

If an agent of degree k chooses action 1 in equilibrium, it follows from degree independence (again, assuming for the sake of the example that n is infinitely large) that an agent of degree $k - 1$ faces a lower likelihood of an arbitrary neighbour choosing the action 1, and would be best responding with action 1 as well. In particular, any equilibrium is characterized by a threshold.

Let t be the smallest integer for which

$$1-\left[1-\sum_{k=1}^{t} Q(k;p)\right]^{t} \geq 1-c. \tag{2}$$

It is easy to check that an equilibrium σ must satisfy $\sigma(k) = 1$ for all $k < t$, $\sigma(k) = 0$ for all $k > t$ and $\sigma(t) \in [0, 1]$. In particular, $\sigma(k)$ is non-increasing.

Observe that social connections create personal advantages: players with degree greater than t obtain higher expected payoffs as compared to the players of degree less than t. In general, the existence and uniqueness of such a symmetric threshold equilibrium follows from simple arguments for binary-action games, both for the present case of strategic substitutes and for the case of strategic complements.[10] For general games, we establish a similar conclusion that every symmetric equilibrium strategy is monotone.

We now look at how equilibrium play is affected by changes in the network. Consider, in particular, a change in the probability distribution over the degrees of players' neighbours that reflects an unambiguous increase in connectivity, as given by the standard criterion of FOSD. Specifically, suppose we move from p to p' where $p' > p$, so that $Q(k; p')$ FOSD $Q(k; p)$. From equation (2), it follows that the (unique) threshold t' corresponding to p' must be higher than t . This has a two-fold implication. First, contingent on any given type, the extent of information acquisition (or public good contribution) does not fall – it remains unchanged for

[10] Naturally, if actions are strategic complements, playing action 1 is prescribed by the equilibrium strategy if the type is no lower than the corresponding threshold. On this issue, see our second example in this section.

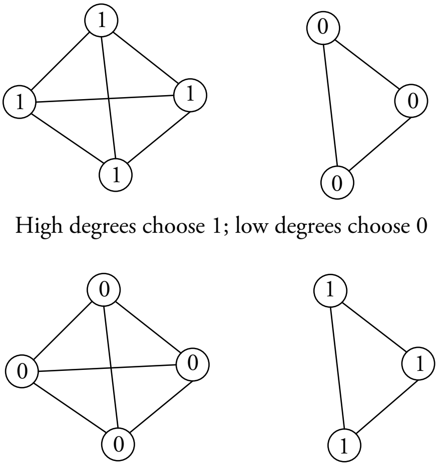
agents with degrees lower than t or greater than t', and increases for all other agents. Second, the probability that any *randomly selected* neighbour of an agent makes a positive contribution falls – for consistency, it must be that $\sum_{k=1}^{t'} Q(k;p') \leq \sum_{k=1}^{t} Q(k;p)$.

This example illustrates the existence of a unique non-increasing symmetric equilibrium, and the two effects of an increase in connectivity: generating a (unique) equilibrium with greater contribution, although reducing the probability that any random neighbour contributes. Our results generalize these insights to a wide array of games exhibiting strategic substitutability, allowing for more general action spaces, payoff structures and neighbour degree correlations. We next study a simple game where actions are strategic complements. Again, consider a context where $X = \{0, 1\}$ is the action space, but now let the payoffs of any particular agent i be given by $(\alpha \bar{x}_{N_i} - c)x_i$ Assuming that $c > \alpha > 0$, these payoffs define a coordination game where, depending on the underlying network and the information conditions, there can generally be multiple equilibria.

As before, we start our discussion with the case of complete information, i.e., with the assumption that the prevailing network is common knowledge. Clearly, the induced game always allows for an equilibrium where $x_i = 0$ for all i. There are generally other equilibria and we illustrate this for a simple network with seven players, split into two complete components with three and four players, respectively. It is easy to see that there is an equilibrium in which all players in the larger component choose 1, whereas all players in the smaller component choose 0. However, the reverse pattern, in which all players in the large component choose 0, whereas all players in the small component choose 1 is also an equilibrium. These are depicted in Figure 3.

By contrast, if we make the assumption that each player is only informed of her own degree (and has independent beliefs on the degrees of neighbours), we find much more definite predictions with regard to equilibrium behaviour. Take, for example, the Erdös–Renyi model and the resulting binomial beliefs considered above. Note that independence of neighbour degrees implies that the probability a random neighbour chooses the action 1 cannot depend on one's own degree. In particular, the expectation of the sum of actions $\bar{x}_{N_i}$ of any agent i with $|N_i| = k$ neighbours is increasing in k. The structure of payoffs then assures that if a degree k agent is choosing the action 1 in equilibrium, any agent of degree greater than k must be best responding with the action 1 as well, and so every equilibrium is determined by a threshold and is non-decreasing. Certainly, everyone choosing the action 0 is a symmetric (threshold) equilibrium. For sufficiently large p there exists $t < N - 1$, an integer, for which

$$\alpha(t-1)\sum_{k=t}^{N-1} Q(k;p) < c \quad \text{and} \quad \alpha t \sum_{k=t}^{N-1} Q(k;p) \geq c.$$



High degrees choose 1; low degrees choose 0

High degrees choose 0; low degrees choose 1

Figure 3. Strategic complements with complete information

Such a t corresponds to an equilibrium that satisfies $\sigma(k) = 0$ for all $k < t$, $\sigma(t) \in [0, 1]$ and $\sigma(k) = 1$ for all $k > t$.

Furthermore, increasing connectedness, as before, by shifting p to p', where $p' > p$, thereby inducing an FOSD shift in neighbours' degree distribution, implies that, $\sum_{k=t}^{N-1} Q(k; p') > \sum_{k=t}^{N-1} Q(k; p)$. Hence, there exists an equilibrium threshold t' corresponding to p' and satisfying $t' \leq t$. Intuitively, the shift to p' increases perceived connectivity and therefore, ceteris paribus, the probability each random neighbour chooses the action 1. Therefore, the value of the action 1 increases, and if all agents use the threshold t, the best response of any particular agent would be to use a threshold lower than t . Continuing such a process iteratively, we generate t'. Note also that $t' \leq t$ implies that the probability that a random neighbor chooses the action 1 in the t' equilibrium under p', given as), $\sum_{k=t'}^{N-1} Q(k; p')$, is greater than the probability that a random neighbour chooses the action 1 in the original equilibrium (with threshold t) under p.

Our results in Section 5 extend these observations to a general class of games with complements: allowing for a wide scope of action spaces, payoff structures and neighbour degree correlations.

To summarize, under complete network information there is no systematic relation between social networks and individual behaviour and payoffs (even if we restrict attention to equilibrium in which players with the same degree choose the same action). By contrast, under incomplete network information, both in games of strategic substitutes as well as in games of strategic complements, we obtain a clear cut relation between networks and individual behaviour and payoffs. Moreover, our discussion clarifies how networks have systematically different effects in games with substitutes and in games with complements.

3. The Model

This section presents our theoretical framework. We start by describing the modelling of a network game, comprised of the degree distribution and each agent's actions and payoffs. We then discuss our equilibrium concept, symmetric Bayesian equilibrium.

3.1. Networks and payoffs

There is a finite set of agents, $N = \{1, 2, ..., n\}$. The connections between them are described by a network that is represented by a matrix $g \in \{0, 1\}^{n \times n}$, with $g_{ij} = 1$ implying that i's payoff is affected by j's behaviour. We follow the convention of setting $g_{ii} = 0$ for all $i \in N$.

Let $N_i(g) = \{j \mid g_{ij} = 1\}$ represent the set of neighbours of i. The *degree* of player i, $k_i(g)$, is the number of i's connections:

$$k_i(g) = |N_i(g)|.$$

Each player i takes an action x_i in X, where X is a compact subset of $[0, 1]$. Without loss of generality, we assume throughout that $0, 1 \in X$. We consider both discrete and connected action sets X. The payoff of player i when the profile of actions is $x = (x_1, ..., x_n)$ is given by:

$$v_{k_i(g)}(x_i, x_{N_i(g)})$$

where $x_{N_i(g)}$ is the vector of actions taken by the neighbours of i. Thus, the payoff of a player depends on her own action and the actions that her neighbours take.

Note that the payoff function depends on the player's degree k_i but not on her identity i. Therefore, any two players i and j who have the same degree ($k_i = k_j$) have the same payoff function. We also assume that v_k depends on the vector $x_{N_i(g)}$ in an anonymous way, so that if x' is a permutation of x (both k-dimensional vectors) then $v_k(x_i, x) = v_k(x_i, x')$ for any x_i. If X is not a discrete set then we assume that it is connected, in which case v_k is taken to be continuous in all its arguments and concave in its own action.

Finally, we turn to the relation between players' strategies and their payoffs. We say that a payoff function exhibits *strategic complements* if it has increasing differences: for all k, $x_i > x_i'$ and $x \geq x'$:

$$v_k(x_i, x) - v_k(x_i', x) \geq v_k(x_i, x') - v_k(x_i', x').$$

Analogously, we say that a payoff function exhibits *strategic substitutes* if it has decreasing differences: for all k, $x_i > x_i'$ and $x \geq x'$:

$$v_k(x_i, x) - v_k(x_i', x) \leq v_k(x_i, x') - v_k(x_i', x').$$

These notions are said to apply strictly if the payoff inequalities are strict whenever $x \neq x'$.

We also keep track of the effects of others' strategies on a player's payoffs. We say that a payoff function exhibits *positive externalities* if for each k, and for all $x \geq x'$, $v_k(x_i, x) \geq v_k(x_i, x')$. Analogously, we say that a payoff function exhibits *negative externalities* if for each k, and for all $x \geq x'$, $v_k(x_i, x) \leq v_k(x_i, x')$. Correspondingly, the payoff function exhibits *strict* externalities (positive or negative) if the above payoff inequalities are strict whenever $x \neq x'$.

We now present some economic examples to illustrate the scope of our framework in terms of the payoff structures it allows for (that will be layered upon the social network configurations we describe below).

EXAMPLE 1. *Payoffs depend on the sum of actions*. Player *i*'s payoff function when she chooses x_i and her k neighbours choose the profile $(x_1, ..., x_k)$ is:

$$v_k(x_i, x_1, ..., x_k) = f\left(x_i + \lambda \sum_{j=1}^{k} x_j\right) - c(x_i), \tag{3}$$

where $f(\cdot)$ is non-decreasing and $c(\cdot)$ is a 'cost' function associated with own effort. The parameter $\lambda \in \mathbb{R}$ determines the nature of the externality across players' actions. This example exhibits (strict) strategic substitutes (complements) if, assuming differentiability, $\lambda f''$ is negative (positive).

The case where f is concave, $\lambda = 1$, and $c(\cdot)$ is increasing and linear corresponds to the case of information sharing as a local public good studied by Bramoullé and Kranton (2007), where actions are strategic substitutes. In contrast, if $\lambda = 1$, but f is convex (with $c'' > f'' > 0$), we obtain a model with strategic complements, which nests a model studied by Goyal and Moraga-Gonzalez (2001) regarding collaboration among firms. In fact, the formulation in equation (3) is general enough to accommodate a good number of further examples in the literature such as human capital investment (Calvó-Armengol and Jackson, 2009), crime networks (Ballester et al., 2006), some coordination problems (Ellison, 1993) and the onset of social unrest (Chwe, 2000).

An interesting special case of Example 1 is the best shot game described in the opening example of Section 2.

EXAMPLE 2 – *'Best shot' public goods games*. The best shot game is a good metaphor for many situations in which there are significant spillovers between players' actions. $X = \{0, 1\}$ and the action 1 can be interpreted as acquiring information (or providing any local and discrete public good). We suppose that $f(0) = 0$, $f(x) = 1$ for all $x \geq 1$, so that acquiring one piece of information suffices. Costs, on the other hand, satisfy $0 = c(0) < c(1) < 1$ so that no individual finds it optimal to dispense with the information but prefers one of her neighbours to gather it. This is a game of strategic substitutes and positive externalities.[11]

In the above examples, a player's payoffs depend on the sum of neighbours' strategies and all of them satisfy the following general property.

Property A. $v_{k+1}(x_i, (x, 0)) = v_k(x_i, x)$ *for any* $(x_i, x) \in X^{k+1}$

Under Property A, adding a link to a neighbour who chooses action 0 is payoff equivalent to not having an additional neighbour. The above discussion clarifies that many economic examples studied so far satisfy Property A. There is however a prominent case where the payoffs violate Property A: this arises when payoffs depend on the average of the neighbours' actions. Our framework allows for of such games as well, like Example 3 below.

EXAMPLE 3 – *Payoffs depend on the average of neighbours' actions*. Let $X = \{0, 1\}$. Player i's payoff function when she chooses x_i and her k neighbours choose the profile $(x_1, ..., x_k)$ is:

$$v_k(x_i, x_1, \ldots, x_k) = x_i f\left(\frac{\sum_{j=1}^{k} x_j}{k}\right) - c(x_i), \tag{4}$$

where $f(\cdot)$ is an increasing function. This is a game of strategic complements and positive externalities.

3.2. Information

We study an environment in which individuals are aware of their proclivity to interact with others, but do not know who these others will be when taking actions. For instance, a researcher choosing an operating system may know the number of coauthors they tend to work with at any given time, but not necessarily who these people will be during the upcoming year. These considerations motivate the informational assumptions in our model: individuals know the number of their contacts and have information on the distribution of connections in the population at large.

Formally, let the degrees of the neighbours of a player i of degree k_i be denoted by $\mathbf{k}_{N(i)}$, which is a vector of dimension k_i. The information a player i of degree k_i has regarding the degrees of her neighbours is captured by a distribution $P(\mathbf{k}_{N(i)} \mid k_i)$. Throughout, we model players' beliefs with a common prior and ex-

[11] For instance, consumers learn from relatives and friends (Feick and Price, 1987), innovations often get transmitted between firms and experimentation is often shared amongst farmers (Foster and Rosenzweig, 1995; Conley and Udry, 2010). For a discussion of best shot games, see Hirshleifer (1983).

ante symmetry. Players may end up with different positions in a network and conditional beliefs, but their beliefs are only updated based on their realized position and not on their names. This means that the information structure is given by a family of anonymous conditional distributions $\mathbf{P} \equiv \{[P(\mathbf{k} \mid k)]_{\mathbf{k}\in\mathbb{N}^k}\}_{k\in\mathbb{N}}$. In some of our results, we also need to refer to the underlying unconditional degree distribution, which is denoted by $P(\cdot)$.

We would like to emphasize that our framework allows for correlation between neighbours' degrees. This means that the conditional distributions concerning neighbours' degrees can in principle vary with a player's degree. This is particularly important in face of the empirical evidence illustrating that social networks generally display such internode correlations. Newman (2003), for example, summarizes empirical results in this respect across different contexts. He reports, specifically, that some networks such as those of scientific collaboration (reflecting joint authorship of papers) or actor collaboration (film co-starring) display significant positive degree correlation while others, such as the internet (physical connections among routers) or the world wide web (hyperlinks between webpages), have a negative one. As these correlations, positive or negative, may well have some bearing (in interplay with game payoffs) on the strategic problem faced by agents, they should be accommodated by the model.

To deal with this issue, we generalize (i.e. weaken) a standard definition of affiliation that has been amply used in the economic literature to capture statistical correlations between collections of random variables (e.g. individual valuations in auctions, as in Milgrom and Weber (1982)).[12] To introduce the notion formally, denote by $\mathbf{k}_{N(i)} = (k_1, k_2, \ldots, k_{k_i})$ the degrees of the neighbours of a typical player with degree . Then, given any function f: $\{0, 1, \ldots, n-1\}^m \rightarrow \mathbb{R}$ where $m \leq k_i$, let

$$E_{P(\cdot|k_i)}[f] = \sum_{\mathbf{k}_{N(i)}} P(\mathbf{k}_{N(i)} \mid k_i) f(k_1, \ldots, k_m). \tag{5}$$

The above expression simply fixes some subset $m \leq k_i$ of i's neighbours, and then takes the expectation of f operating on their degrees. We say that $\mathbf{P}$ exhibits *positive neighbour affiliation* if, for all $k' > k$, and any non-decreasing f: $\{0, 1, \ldots, n-1\}^k \rightarrow \mathbb{R}$.

$$E_{P(\cdot|k')}[f] \geq E_{P(\cdot|k)}[f]. \tag{6}$$

Analogously, $\mathbf{P}$ exhibits *negative neighbour affiliation* if the reverse inequality holds for each $k' > k$ and non-decreasing f.

[12] Affiliation, in turn, can be viewed as a strengthening of the notion of association that is common in the statistical literature – see, e.g. Esary et al.(1967) for a useful reference. In this paper, we are interested in both the notion of positive affiliation (which is the usual case postulated in the literature) as well as a negative one – the conditions and implications, however, are obviously fully symmetric in each case.

As indicated, our notion of neighbour affiliation is weaker than what affiliation (positive or negative) among the whole vector of random variables (k_i, $\mathbf{k}_{N(i)}$) would entail.[13] It simply embodies the idea that higher degrees for a given player are correlated with higher or lower degree (depending on whether it is positive or negative, respectively) of *all* her neighbours. Obviously, it is satisfied in the case where neighbours' degrees are all stochastically independent. This is, for example, a condition that holds asymptotically in many models of random networks, including the classical model of Erdös–Rényi or the more recent configuration model (see, e.g. Newman, 2003; Vega-Redondo, 2007; and Jackson, 2008 for discussions). Positive neighbour affiliation, on the other hand, is a feature commonly found in other models of network formation that have a dynamic dimension – cf. the model of Barabàsi and Albert (1999) based on preferential attachment or the models by Vazquez (2003) and Jackson and Rogers (2007) reflecting network-based search.[14] In addition, an important motivation for internode degree correlations is empirical. For, as mentioned, many of the studies on real social networks undertaken in recent years find strong evidence for either positive or negative correlations. Neighbour affiliation, while entailing some restrictions, provides a workable tool for capturing these observations.

Finally, we also need a way of comparing situations where the network (and thus the corresponding beliefs) undergo changes in connectivity. We focus on changes that reflect unambiguous increases or decreases in the distribution of agents' degrees. So we use a suitable extension of the standard notion of FOSD to embody changes in the degree distributions that capture the idea of link addition. Specifically, we say that $\mathbf{P}'$ *dominates* $\mathbf{P}$ if for all k, and any non-decreasing f: $\{0, 1, ..., n - 1\}^k \to \mathbb{R}$

$$E_{P'(\cdot|k)}[f] \geq E_{P(\cdot|k)}[f].$$

This concept of dominance is a generalization of stochastic dominance relationships adapted to vectors and families of distributions.

To conclude, a network game is fully described and is henceforth denoted by a quadruple (N, X, $\{v_k\}_k$, $\mathbf{P}$). In certain cases, concentrating on degree distributions that exhibit independence between neighbours' degrees allows us to derive further insights. In these cases, the entire set of conditionals is captured by the underlying distribution P and so we denote the corresponding network game by (N, X, $\{v_k\}_k$, P).

[13] To see this, refer to Theorem 5 in Milgrom and Weber (1982), which establishes that affiliation implies that the counterpart of equation (6) must hold when we condition on any subset of the random variables in (k_i, $\mathbf{k}_{N(i)}$) and compute the expected value for any non-decreasing function of those random variables. More precisely, our notion of neighbour affiliation is identical to the concept of *positive regression dependence* with respect to k_i, as formulated by Lehmann (1966). Esary et al. (1967) show that this concept is weaker than the standard one of association, except for bivariate random variables.

[14] See the working paper version, (Galeotti et al. 2006) for a formal description of neighbour affiliation attributes of commonly observed and studied network formation procedures.

3.3. The Bayesian game

A *strategy* for player i is a mapping $\sigma_i : \{0, 1, ..., n-1\} \rightarrow \Delta(X)$, where $\Delta(X)$ is the set of probability distributions on X. So, $\sigma_i(k)$ is the mixed strategy played by a player of degree k. We analyse (symmetric) Bayesian equilibria of this game and they can be represented simply as a (mixed) strategy, $\sigma(\cdot)$.[15]

More formally, given a player i of degree k_i let $\psi(x_{N_i(g)}, \sigma, k_i)$ be the probability distribution over $x_{N_i(g)} \in X^{k_i}$ induced by the beliefs $P(\cdot| k_i)$ over the degrees of i's neighbours when composed with the strategy σ. Thus, the expected payoff to a player i with degree k_i when other players use strategy σ and i chooses action x_i is

$$U(x_i,\sigma,k_i) \equiv \int_{x_{N_i(g)} \in X^{k_i}} v_{k_i}(x_i, x_{N_i(g)}) d\psi(x_{N_i(g)}, \sigma, k_i). \tag{7}$$

A strategy σ comprises a symmetric *Bayesian equilibrium* (or just an equilibrium, for short) if $\sigma(k_i)$ is a best response, for each degree k_i to the strategy σ being played by other players. That is, σ is an equilibrium if for every degree ki displayed by any typical agent i, the following holds:

$$U(x_i,\sigma, k_i) \geq U(x_i',\sigma, k_i), \ \forall \ x_i' \in X, x_i \in \mathbf{supp}(\sigma(k_i)). \tag{8}$$

Our interest is in understanding the effects of networks on behaviour and welfare. To bring out these effects clearly, we focus on symmetric Bayes-Nash equilibria, i.e. configurations where all players with the same network characteristic (which, under our informational assumptions, is their degree) choose the same strategy. This is further motivated by the observation that, in fact, *all* equilibria of the game must be symmetric when the following two conditions apply:

(i) the underlying network formation mechanism is anonymous and the population very large;
(ii) the payoff function is strictly concave in own action.

For, in this case, all agents of any given degree face the same decision problem (from (i)) and the optimal choice in it is unique (by (ii)). This leads to a symmetric behaviour.[16]

[15] Static equilibrium refinements are not so useful in our case, as our equilibria are typically strict; e.g., in our applications (as, say, in best shot games of the sort discussed in Section 2), both in the complete and incomplete information scenarios. Finally, it is worth noting, refinements that require dynamic stability in terms of an adjustment process can encounter non-existence problems. As an illustration, consider the notion of stable equilibrium used by Bramoullé and Kranton (2007) for their analysis of local public goods in networks. As they show, these equilibria exist only for networks whose maximal independent set has two nodes in every non-provider's neighbourhood, which rules out many networks.

[16] Formally, the statement here is in effect of an asymptotic nature, pertaining to the limit equilibrium behaviour as the population size grows infinite. To be precise, consider the relatively simple case where the

It is worth emphasizing as well that the contrast between complete and incomplete information that is the heart of our analysis remains in force when we restrict attention to symmetric equilibria in both cases. To illustrate this, recall the star networks that were considered in Section 2 (see especially Figure 1). There, the restriction of symmetry under complete information requires that all peripheral players choose the same action. But, as we saw, this allows for two polar and very different Nash equilibria. Instead, symmetry under incomplete information singles out a unique equilibrium outcome in which the centre does not contribute. Our discussion in Section 2 suggests that analogous observations hold for games with strategic complements (see Figure 3).

To relate network structure and the primitives of the payoffs to features of equilibrium, we need to relate strategies to degrees. Some basic definitions of monotonicity are thus useful in stating our results.

A strategy σ is *non-decreasing* if $\sigma(k_i)$ first-order stochastically dominates $\sigma(k)$ for each $k' > k$. Similarly, σ is *non-increasing* if the domination relationship is reversed.

Expected payoffs exhibit *degree complementarity* if

$$U(x_i, \sigma, k_i) - U(x_i', \sigma, k_i) \geq U(x_i, \sigma, k_i') - U(x_i', \sigma, k_i'),$$

whenever $x_i > x_i'$, $k_i > k_i'$ and σ is non-decreasing. Analogously, payoffs exhibit *degree substitution* if the inequality above is reversed in the case where σ is non-increasing.

Degree complementarity captures the idea that if a high strategy is more attractive than a low strategy for a player of some degree, then the same is true for a player of a higher degree when the strategy being played by other players is non-decreasing. Degree complementarity arises in many contexts that are covered by our framework. We illustrate this by considering two cases of interest.

Recall that Property A says that $v_{k+1}(x_i, (x, 0)) = v_k(x_i, x)$ for any $(x_i, \mathrm{x}) \in X^{k+1}$. We note that Property A, strategic complements of $v_k(\cdot, \cdot)$ and positive neighbour affiliation of P ensure degree complementarity. To see why this is true consider a strategy σ which is non-decreasing and suppose that $k' = k + 1$. Now observe that

$$\begin{aligned}
&U(x_i, \sigma, k) - U(x_i', \sigma, k) \\
&\quad = \int_{x \in X^k} [v_k(x_i, x) - v_k(x_i', x)] d\psi(x, \sigma, k) \\
&\quad = \int_{x \in X^k} [v_{k'}(x_i, (x, 0)) - v_{k'}(x_i', (x, 0))]\, d\psi(x, \sigma, k) \\
&\quad \leq \int_{x \in X^k} [v_{k'}(x_i, (x, 0)) - v_{k'}(x_i', (x, 0))]\, d\psi((x, 0), \sigma, k') \\
&\quad \leq \int_{(x, x_{k+1}) \in X^{k'}} [v_{k'}(x_i, (x, x_{k+1})) - v_{k'}(x_i', (x, x_{k+1}))]\, d\psi((x, x_{k+1}), \sigma, k') \\
&\quad = U(x_i, \sigma, k') - U(x_i', \sigma, k')
\end{aligned}$$

underlying network formation mechanism is random, the degree distribution has a uniformly bounded support and every two networks differing only in some arbitrary permutation of player indices have an identical ex-ante probability. Under these conditions, the probability that any two agents be connected becomes insignificant for large populations and, therefore, if they have the same degree, they must also face a probability distribution over neighbours' actions that is essentially the same. Then, by strict concavity and continuity of payoffs, the claim follows.

where the second equality follows from Property A, the first inequality follows from positive neighbour affiliation, σ being non-decreasing and strategic complements, while the second inequality follows from strategic complements. Analogous considerations establish that Property A, strategic substitutes of $v_k(\cdot,\cdot)$ and negative neighbour affiliation of $\mathbf{P}$ ensure degree substitution.

While Property A (taken along with the corresponding properties on $\mathbf{P}$ and $v_k(\cdot,\cdot)$) is sufficient to establish degree complementarity and substitution, it is not necessary. The following discussion, which builds on Example 3, illustrates this point.

EXAMPLE 4 – *Degree complements and substitutes without Property A.* Suppose that payoffs are as in Example 3. In addition, let $\mathbf{P}$ be such that neighbours' degrees are stochastically independent (e.g. as in an asymptotic Erdös–Rényi random network discussed in Section 2). When neighbours' degrees are independent, $\frac{kP(k)}{\langle k\rangle}$ captures the probability that a random neighbour is of degree k (see, e.g. Jackson, 2008). Let Y_m be a random variable that has a binomial distribution with m draws each with probability $\sum_k \frac{kP(k)}{\langle k\rangle}\sigma(k)$, the expected action of any neighbour. Then, the expected payoffs to a player i are given by:

$$U(x_i,\sigma,k_i)=E\left[x_i f\left(\frac{Y_{k_i}}{k_i}\right)\right]-c(x_i),$$

and thus

$$U(1,\sigma,k_i)-U(0,\sigma,k_i)=E\left[f\left(\frac{Y_{k_i}}{k_i}\right)\right]-c(1)+c(0).$$

Note that $\frac{Y_{k'}}{k'}$ is a mean preserving spread of $\frac{Y_k}{k}$ when $k' < k$. Thus, if f is concave, we have degree complementarity, whereas if f is convex then degree substitution obtains.

4. Equilibrium: Existence and Monotonicity

We start by showing existence of an equilibrium involving monotone strategies. We then provide conditions under which all equilibria are monotone. Finally, we close the section by exploring the relationship between network degree and equilibrium payoffs. The latter analysis, in particular, identifies conditions under which payoffs increase/decrease with network degree, thereby clarifying the contexts in which network connections are advantageous and disadvantageous, respectively.

Recall that a strategy σ is *non-decreasing* if $\sigma(k')$ first-order stochastically dominates $\sigma(k)$ for each $k' > k$. Similarly, σ is *non-increasing* if the domination relationship is reversed.

Proposition 1. *There exists a symmetric equilibrium, and if the game has degree complements, then there exists a symmetric equilibrium in pure strategies. If there is degree complementarity (substitution) then there is a symmetric equilibrium that is non-decreasing (non-increasing).*

To show the validity of this result, we start by addressing the existence of a symmetric equilibrium. It has been assumed that players have identical action sets X, the payoff functions are also the same and player's beliefs concerning network are ex-ante symmetric. The game, therefore, is a symmetric one of incomplete information. Given that the action set is compact, the payoff function is continuous in all arguments (when the action set is non-discrete) and concave in own action, it is then straightforward to adapt the usual fixed-point argument to show that there exists a symmetric equilibrium, possibly in mixed strategies. Moreover, the fact that this symmetric equilibrium can be chosen in pure strategies under degree complements follows from standard strategic complements arguments (see, e.g. Milgrom and Shannon, 1994).

On the other hand, concerning monotonicity, one can readily exploit the degree complements/substitutes property to show that for a player faced with a monotone strategy played by the rest of the population, there always exists a monotone best reply. Then, as the set of monotone strategies is convex and compact, the existence of a monotone equilibrium derives from standard arguments (see, e.g. Milgrom and Shannon, 1994; van Zandt and Vives, 2007).

Next, we elaborate on two aspects of Proposition 1. *First*, we discuss whether *every* symmetric equilibrium is monotone. Consider a game with action set $X = \{0, 1\}$ and payoffs $v_k(x_i, x_{N_i(g)}) = x_i \sum_{j \in N_i(g)} x_j - c\, x_i$, where $0 < c < 1$ (a special case of the second example in Section 2). This example satisfies Property A and the underlying game displays strategic complements. Now suppose that there is perfect degree correlation so that players are connected to others of the same degree. It is then clear that *any* symmetric pure strategy profile defines an equilibrium.[17] This example suggests that the possibility of non-monotone equilibria is related to the correlation in degrees. This point is highlighted by the following result.

Proposition 2. *Suppose that payoffs satisfy Property A and that the degrees of neighbouring nodes are independent. Then, under strict strategic complements (substitutes) every symmetric equilibrium is non-decreasing (non-increasing).*

The key point to note here is that, under independence, degree k and degree $k' = k + 1$ players have the same beliefs about the degree of each of their

[17] In fact, the best response of a degree k player is to choose 0 (1) if all other degree k players also choose 0 (1).

neighbours. If the $k + 1$th neighbour is choosing 0 then under Property A the degree k' player will choose the same best response as the degree k player; if the $k + 1$th neighbour chooses a positive action then strict complementarities imply that the degree k' player best responds with a higher action.[18]

Going back to some of our motivating examples, Proposition 2 has very clear implications. Consider the student aiming at a career of diplomacy and contemplating learning a new language. As the value of knowing a language is increasing with the number of connected individuals who speak that language (it is a game of complements), we would expect that student to be more likely to take on the study of the new language than a student who aims at a less interactive career.

A *second* issue is whether the nature of degree correlation – positive neighbour affiliation under strategic complements or negative neighbour affiliation under strategic substitutes – is essential for existence of monotone equilibria. Consider a special case of Example 1 in which $X = [0, 1]$, $f(y) = \gamma y + \alpha y^2$, $y = x_i + \sum_{j \in N_i(g)} x_j$ and $c(x_i) = \beta x_i^2$ for some $\gamma, \alpha, \beta > 0$. This game exhibits strategic complements. Next suppose that the unconditional degree distribution satisfies $P(1) = P(2) = \varepsilon$ and $P(\bar{k}) = 1 - 2\varepsilon$ for some small ε and a given large $\bar{k}$. Further suppose that $P(\bar{k}\,|1) = P(2|2) = 1$, i.e., all agents with degree 1 are connected to those of degree $\bar{k}$ and all those of degree 2 are connected among themselves. Note that this pattern of connections violates positive neighbour affiliation. It is now possible to verify that if $\beta > \alpha$ then *every* equilibrium is interior; moreover if $\bar{k}$ is large enough and ε sufficiently small then σ satisfies $\sigma_2 < \sigma_1 < \sigma_{\bar{k}}$ and is not monotone.

A recurring theme in the study of social structure in economics is the idea that social connections create personal advantages. In our framework, the relation between degrees and payoffs is the natural way to study network advantages. Let us consider games with positive externalities and positive neighbour affiliation, and look at a player with degree $k + 1$. Suppose that all of her neighbours follow the monotone increasing equilibrium strategy, but her $k + 1$th neighbour chooses the minimal 0 action. Property A implies that our $(k + 1)$ degree player can assure herself an expected payoff which is at least as high as that of any k degree player by simply using the strategy of the degree k player. These considerations lead us to state the following result.

***Proposition 3**. Suppose that payoffs satisfy Property A. If* P *exhibits positive neighbour affiliation and the game displays positive externalities (negative externalities), then in every non- decreasing symmetric equilibrium the expected payoffs are non-decreasing (non-increasing) in degree. If* P *exhibits negative neighbour affiliation and the game displays*

[18] The strictness is important for the result. For instance, if players were completely indifferent between all actions, then non-monotone equilibria are clearly possible.

positive externalities (negative externalities), then in every non-increasing symmetric equilibrium the expected payoffs are non-decreasing (non-increasing) in degree.

We emphasize that under positive externalities, players with more neighbours earn higher payoffs irrespective of whether the game exhibits strategic complements or substitutes (under the appropriate monotone equilibrium). These network advantages are especially striking in games with strategic substitutes (such as local public goods games) and negative neighbour affiliation: here higher degree players exert lower efforts but earn a higher payoff as compared to their less connected peers.

5. The Effects of Changing Networks

We now investigate how changes in a network – such as the addition/deletion of links or the redistribution of links away from a regular network to highly unequal distributions that characterize empirically observed networks – affect the behaviour and welfare of players. We start with games of strategic substitutes and then take up games of strategic complements.

5.1. Games with strategic substitutes

We refer to games where payoffs are of strict strategic substitutes and satisfy Property A and where $\mathbf{P}$ exhibits negative neighbour affiliation as *binary network games of substitutes*, and we focus on such games in the following analysis. An attractive feature of binary action network games with substitutes is that there is a unique symmetric equilibrium strategy σ, and it involves a threshold.

Proposition 4. *Consider a binary network game of substitutes. There exists some threshold $t \in \{0, 1, 2,.. .\}$ such that the probability $\sigma(1|\cdot)$ of choosing action 1 in the unique non-increasing symmetric equilibrium strategy σ satisfies $\sigma(1|\, k_i) = 1$ for $k_i < t$, $\sigma(1|\, k_i) = 0$ for all $k_i > t$ and $\sigma(1|t) \in (0, 1]$ for $k_i = t$.*

Now we ask: what is the effect of adding links on equilibrium behaviour? We first observe that the best response of a player depends on the actions and hence the expectations concerning the degrees of her neighbours. Thus, the effects of link addition must be studied in terms of the change in the degree distribution of the neighbours.[19] We therefore approach the addition of links in terms of an

[19] Indeed, it is important to note that the relationship between two underlying (unconditional) degree distributions does not imply a similar relation for the conditional distribution over neighbours' degrees, even under independence. As an illustration consider a case where degrees of neighbours are independent. Consider two degree distributions P and P', where P' assigns one half probability to degrees 2 and 10 each, whereas distribution P assigns one half probability to degrees 8 or 10 each. Clearly P FOSD P'. As mentioned above, when neighbouring degrees are independent, the probability of having a link with a node is (at least roughly, depending on the process) proportional to the degree of that node, so that for all

increase in the degrees of a neighbour. In our context of non-increasing strategies, this means a fall in her action (on average), which, from strategic substitutes, suggests that the best response of the player in question should increase. However, this increase in action of every degree may come into conflict with the expectation that neighbours must be choosing a lower action, on average. The following result clarifies how this tension is resolved. Denote by t the threshold in the game $(N, X, \{v_k\}_k, \mathrm{P})$ and by t' the threshold in game $(N, X, \{v_k\}_k, \mathrm{P}')$.

Proposition 5. *Let* $(N, X, \{v_k\}_k, \mathrm{P})$ *and* $(N, X, \{v_k\}_k, \mathrm{P}')$ *be binary network games of substitutes. If* P *dominates* P'*, then* $t \geq t'$*. However, for the threshold degree type* t *the probability that a neighbour chooses 1 is lower under* P.

This result clarifies that an increase in threshold for choosing 1 is consistent with equilibrium behaviour because each of the neighbours is more connected and chooses 1 with a lower probability (in spite of an increase in the threshold). The best shot game helps to illustrate the effects of dominance shifts in degrees which are derived in the above result.

EXAMPLE 5 – *Effects of increasing degrees in a best shot game*. Consider the best shot game discussed in the introduction and described in Example 2. Set $c = 25/64$. Suppose that degrees take on values 1, 2 and 3 and that the degrees of neighbours are independent. Note that, in view of Proposition 2 and Proposition 4, the assumption that the degrees of neighbouring nodes are stochastically independent implies that there exists a unique symmetric equilibrium which is non-increasing and it is fully characterized by a threshold.

Let us start with initial beliefs P' that assign probability one-half to neighbouring players having degrees 1 and 2. In the unique symmetric equilibrium, degree 1 players choose 1 with probability 1, whereas degree 2 players choose 1 with probability 0. Hence, at equilibrium, the probability that a neighbour of a degree 2 player chooses action 1 is ½.

Consider now a dominance shift to P, so that neighbouring players are believed to have degrees 2 and 3 with probability one-half each. It can be checked that the unique equilibrium involves degree 2 players choosing action 1 with probability ¾, whereas degree 3 players choose 1 with probability 0. Consequently, the probability that a neighbour of a degree 2 player chooses action 1 is 3/8.

Overall, the dominance shift in the beliefs from P' to P leads to an increase in the threshold from 1 to 2. However, the threshold degree 2 player has lower expectation of action 1 under P as compared to P'.

k, $P(k'|k) = k' \, P(k')/\sum P(l)l$. Let $\tilde{P}(k')$ be the neighbour's degree distribution. Under $\tilde{P}'$, the probability that a neighbour has degree 10 is 5/6, while under $\tilde{P}$, the same probability is 5/9. Thus, $\tilde{P}$does not FOSD $\tilde{P}'$.

We now turn to the effects on welfare. The expected welfare is assessed by the expected payoff of a randomly chosen player (according to the prevailing degree distribution). Observe that dominance shifts in the interaction structure lower the expected probability that a randomly selected neighbour of a t-degree player (the threshold player under $\mathbf{P}$) chooses 1. If the degrees of neighbours are independent, then the average effort of a randomly selected neighbour of a player i does not depend on i's degree, and therefore all players expect lower action from each of their neighbours. However, in the presence of negative neighbour affiliation, matters are more complicated, and it is possible that the overall effect of a dominance shift in the distribution of connections can be positive for some degrees and negative for others.

Proposition 5 compares behaviour across networks when there is an increase in the density of links in the sense of domination. However, there are many cases where we might be interested in comparing networks when there is not a clear cut domination relation. We now develop a result on the effect of *arbitrary* changes in the degree distribution.

For simplicity, we focus on the case where degrees of neighbours are independent. Let $\mathbf{P}$ and $\mathbf{P}'$ be two different sets of beliefs and suppose that $\tilde{F}$ and $\tilde{F}'$ are the corresponding induced cumulative distribution functions of the degree distributions, respectively. Let t and t' stand for the threshold types defining the (unique) threshold equilibria under $\mathbf{P}$ and $\mathbf{P}'$, respectively.

Proposition 6. *Let $(N, X, \{v_k\}_k, P)$ and $(N, X, \{v_k\}_k, P')$ be binary network games of substitutes with independent neighbour degrees. Let t and t' denote the unique equilibrium thresholds for these games. If $\tilde{F}\ (t') \leq\ \tilde{F}'\ (t' - 1)$ then $t \geq t'$. Moreover, in these equilibria, the probability that any given neighbour chooses 1 in $(N, X, \{v_k\}_k, P)$ is lower than in $(N, X, \{v_k\}_k, P')$.*

The key issue here is the change in the probability mass relative to the threshold. If the probability of degrees equal or below the threshold goes down then the probability of action 1 decreases and from strategic substitutes, the best response of threshold type t must still be 1. In other words, the threshold rises weakly.

The contribution of Proposition 6 is that it allows us to examine the effect of *any change of the degree distribution*. A natural and important example of such changes is increasing the polarization of the degree distribution by shifting weights to the ends of the support of the degree distribution, as is done under a mean preserving spread (MPS) of the degree distribution. In particular, the above results can be directly applied to the case of strong MPS shifts in the degree distributions. Focusing on the unconditional beliefs (taken to coincide with the unconditional degree distributions because of independence), we say that $P(\cdot)$ is a *strong MPS* of

$P'(\cdot)$ if they have the same mean and there exists L and H such that $P(k) \geq P'(k)$ if $k < L$ or $k > H$, and $P(k) \leq P'(k)$ otherwise. Proposition 6 implies that, in the context of binary-action games, the equilibrium effects of any such change can be inferred from the relative values of the threshold t, L and H.

5.2. Games with strategic complements

This section studies the effects of changes in the network on equilibrium behaviour and payoffs in games with strategic complements. From Proposition 1 we know that equilibria are increasing in degree in games with degree complementarities. As we shift weight to higher degree neighbours each player's highest best response to the original equilibrium profile would be at least as high as the supremum of her original strategy's support. We can now iterate this best response procedure. As the action set is compact, this process converges and it is easy to see that the limit is a (symmetric) non-decreasing equilibrium that dominates the original one. The following result summarizes this argument.

We refer to a network game in which payoffs satisfy strict strategic complements and Property A and P exhibits positive neighbour affiliation as a *network game of complements*.

Proposition 7. *Let* $(N, X, \{v_k\}_k, \mathrm{P})$ *and* $(N, X, \{v_k\}_k, \mathrm{P}')$ *be network games of complements. If* P *dominates* P'*, then for every non-decreasing equilibrium* σ' *of* $(N, X, \{v_k\}_k, \mathrm{P}')$*, there exists a non-decreasing equilibrium* σ *of* $(N, X, \{v_k\}_k, \mathrm{P})$ *that dominates it.*

The proof is straightforward and omitted.[20] Consider next the effect of a dominance shift in the social network on welfare. Recall that the expected welfare is assessed by the expected payoff of a randomly chosen player. Naturally, it must depend on whether the externalities are positive or negative. Suppose, for concreteness, that they are positive and let P dominate P'. Then, from Proposition 7, we know that for every non-decreasing equilibrium σ' under P' there exists a non-decreasing equilibrium σ under P in which players' actions are all at least as high. Hence, the expected payoff of each player is higher under P. However, as expected payoffs are non-decreasing in the degree of a player, to assess welfare it is also important to consider the relation between the corresponding unconditional degree distributions $P(\cdot)$ and $P'(\cdot)$. If, for example, $P(\cdot)$ FOSD $P'(\cdot)$, then, the above considerations imply that the ex-ante expected payoff of a randomly chosen player must rise when one moves from P' to P. We summarize this argument in the following result. For a non-decreasing strategy profile σ

[20] Note that if $(N, X, \{v_k\}_k, \mathrm{P}')$ is a network game of complements, and P dominates P', it is not necessarily the case that P exhibits positive neighbour affiliation. In that case, we can still use similar arguments to construct a new symmetric equilibrium (under P) that dominates σ, although it need not be non-decreasing.

under P, define $W_{\mathrm{P}}(\sigma)$ to be the expected payoff of a node picked at random (under $P(\cdot)$).

Proposition 8. *Let $(N, X, \{v_k\}_k, \mathrm{P})$ and $(N, X, \{v_k\}_k, \mathrm{P}')$ be network games of complements, in which payoffs satisfy positive externalities. Suppose that P dominates P' and $P(\cdot)$ FOSD $P'(\cdot)$. For any non-decreasing equilibrium σ' of $(N, X, \{v_k\}_k, \mathrm{P}')$, there exists a non-decreasing equilibrium σ of $(N, X, \{v_k\}_k, \mathrm{P})$ such that $W_{\mathrm{P}}(\sigma) \geq W_{\mathrm{P}'}(\sigma')$.*

The proof follows from the arguments above and is omitted.

Propositions 7 and 8 pertain to dominance shifts in the conditional degree distributions. However, as in the case of games of substitutes, in binary games with independent degree distributions, we can identify the effects of arbitrary changes in the degree distribution. Indeed, in those games, an analogue of Proposition 4 can be readily established and symmetric equilibria take the form of threshold equilibria: $\sigma(1|k_i) = 0$ for $k_i < t$, $\sigma(1|k_i) = 1$ for all $k_i > t$ and $\sigma(1|t) \in (0, 1]$ for $k_i = t$. Recalling that for any two distributions P and P', $\tilde{F}$ and $\tilde{F}'$ denote their respective cumulitive distributions, we have:

Proposition 9. *Let $(N, X, \{v_k\}_k, P)$ and $(N, X, \{v_k\}_k, P')$ be binary network games of complements and independent neighbour degrees. Let t' be an equilibrium threshold of $(N, X, \{v_k\}_k, P')$.* If $\tilde{F}(t') \leq \tilde{F}'(t'-1)$ *then there is an equilibrium of $(N, X, \{v_k\}, P)$ with corresponding threshold type $t \leq t'$. Moreover, the probability that any given neighbour chooses 1 rises.*

The proof for this result follows along the lines of the proof of Proposition 6 and is omitted.

We conclude by observing that the strategic structure of payoffs has an important effect: recall from Subsection 5.1 that in the case of strategic substitutes, the probability that any neighbour chooses 1 falls when network connectivity grows. By contrast, in games of strategic complements the addition of links leads to an increase in the probability that a neighbour chooses action 1.

6. Deeper Network Information

So far we have focused on the case where players only know their own degree and best respond to the anticipated actions of their neighbours based on the (conditional) degree distributions. We now investigate the implication of increasing the information that players possess about their local networks. As a natural first step along these lines, we examine situations where a player knows not only how many neighbours she has but also how many neighbours each of her neighbours has (e.g. a researcher deciding on an operating system may know the identities of her coauthors, and the number of coauthors that they each have, but nothing beyond that). The arguments we develop in this section extend in a natural way to general radii of local knowledge. Indeed, in the limit, as this radius of knowledge grows, we arrive at complete knowledge of the arrangement of degrees in the network.[21]

Formally, the common type space $\mathcal{T}$ of every player i consists of elements of the form $(k; \ell_1, \ell_2, ..., \ell_k)$ where $k \in \{0, 1, 2, ..., n-1\}$ is the degree of the player and ℓ_j is the degree of neighbour j $(j = 1, 2, ..., k)$, where (in an anonymous setup where the identity of neighbours is ignored) we may assume without loss of generality that neighbours are indexed according to decreasing degree (i.e. $\ell_j \geq \ell_{j+1}$). Given the multi-dimensionality of types in this case, the question arises as to how one should define monotonicity. In particular, the issue is what should be the order relationship $\succeq$ on the type space underlying the requirement of monotonicity. For the case of strategic complements, it is natural to say that two different types, $t = (k; \ell_1, \ell_2, ..., \ell_k)$ and $t' = (k'; \ell'_1, \ell'_2, ..., \ell'_{k'})$, satisfy $t \succeq t'$ if and only if $k \geq k'$ and $\ell_u \geq \ell'_u$ for all $u = 1, 2, ..., k'$. On the other hand, for the case of strategic substitutes, we write $t \succeq t'$ if and only if $k \geq k'$ and $\ell_u \leq \ell'_u$ for all $u = 1, 2, 3, ..., k'$. Given any such (partial) order on $\mathcal{T}$, we say that a strategy σ is non-decreasing if for all $t_i, t_i' \in \mathcal{T}$, $t_i \succeq t_i' \Rightarrow \sigma(t_i)$ FOSD $\sigma(t_i')$. The notion of a non-increasing strategy is defined analogously.

We first illustrate the impact of richer knowledge on the nature of equilibria. It is easier to see the effects of deeper network information in the simpler setting where the degrees of the neighbours are independent and so, for expositional simplicity, we assume independence of neighbours' degrees in this section.[22] Recall from Proposition 2 that under degree independence all symmetric equilibria are non-decreasing (non-increasing) in the case of strategic complements (substitutes) when agents are only informed of their own degree. The following example shows that greater network knowledge introduces non-monotone equilibrium even if the degrees of neighbouring nodes are stochastically independent.

EXAMPLE 6 – *Non-monotone equilibria with knowledge of neighbours' degrees*. Consider a setting where nodes have either degree 1 or degree 2, as given by the corresponding probabilities $P(1)$ and $P(2)$. Suppose that the game is binary action with $X = \{0, 1\}$ and displays strategic complements. Specifically, suppose that the payoff of a player only depends on his own action x_i and the sum $\bar{x}$ of his neighbours' actions as given by a function $v(x_i, \bar{x})$ as follows: $v(0, 0) = 0$, $v(0, 1) = ½$, $v(0, 2) = ¾$, $v(1, 0) = -1$, $v(1, 1) = 1$, $v(1, 2) = 3$.

[21] For results on this limit case, see the earlier version of this paper (Galeotti et al., 2006). See also Kets (2007) for more discussion about the structure of information and its effects.

[22] We note that the assumption of stochastic independence of the degrees of neighbouring nodes implies independence of degrees of neighbours.

It is readily seen that, for any P with support on degrees 1 and 2, the following strategy σ defines a symmetric equilibrium: $\sigma(1|1; 1) = 1$; $\sigma(1|1; 2) = 0$; $\sigma(1|2; \ell_1, \ell_2) = 0$ for any $\ell_1, \ell_2 \in \{1, 2\}$. Here, two players that are only linked to each other both play 1, while all other players choose 0.

Similar non-monotonic equilibrium examples can be constructed for games with strategic substitutes. These observations leave open the issue of whether there exist *any* suitably increasing or decreasing equilibria. The following result shows that a monotone equilibrium always exists if players have deeper network information.

Proposition 10. *Suppose that neighbours' degrees are independent, players know their own degree and the degrees of their neighbours and payoffs satisfy Property A. Under strategic complements (strategic substitutes) there exists a symmetric equilibrium that is non-decreasing (non-increasing).*

The proof of the proposition, which appears in the Appendix, extends naturally the ideas mentioned for the proof of Proposition 2, i.e., the best reply to a monotone strategy can be chosen monotone and the set of all monotone strategies is compact and convex. A direct implication of the result is that there is always an equilibrium that, on average across the types $(k; \ell_1, \ell_2, ..., \ell_k)$ consistent with each degree k, prescribes an (average) action that is monotone in degree. Equipped with the above monotonicity result, it is also possible to recover most of the insights obtained earlier under the assumption that players only know their own degree.

7. Concluding Remarks

Empirical work suggests that the patterns of social interaction have an important influence on economic outcomes. These interaction effects have however been resistant to systematic theoretical study: even in the simplest examples games on networks have multiple equilibria that possess very different properties. The principal innovation of our paper is the introduction of the idea that players have incomplete network knowledge. In particular, we focus on an easily measurable aspect of networks, the number of personal connections/degree, and suppose that players know their own degree but have incomplete information concerning the degree of others in the network. This formulation allows us to develop a general framework for the study of games played on networks. On the one hand, it allows us to accommodate a large class of games with strategic complements and strategic substitutes. On the other hand, it allows us to capture features displayed by real world networks such as general patterns of correlations across the degrees of neighbours.

The analysis of this framework yields a number of powerful and intuitively appealing insights with regard to the effects of location within a network as well as with regard to changes in networks on equilibrium actions and payoffs. These results also clarify how the basic strategic features of the game– as manifest in the substitutes and complements property– combine with different patterns of degree correlations to shape behaviour and payoffs.

In this paper, we have focused on the degree distribution in a network. The research on social networks has identified a number of other important aspects of networks, such as clustering, centrality and proximity, and in future work it would be interesting to bring them into the model.

Appendix

Proof of Proposition 2. We present the proof for the case of strategic complements. The proof for the case of strategic substitutes is analogous and omitted. Let $\{\sigma_k^*\}$ be the strategy played in a symmetric equilibrium of the network game. If $\{\sigma_k^*\}$ is a trivial strategy with all degrees choosing action 0 with probability 1, the claim follows directly. Therefore, from now on, we assume that the equilibrium strategy is non-trivial and that there is some k' and some $x' > 0$ such that $x' \in \mathbf{supp}(\sigma_{k'}^*)$.

Consider any $k \in \{0, 1, \ldots , n\}$ and let $x_k = \sup[\mathbf{supp}(\sigma_k^*)]$. If $x_k = 0$, it trivially follows that $x_{k'} \geq x_k$ for all $x_{k'} \in \mathbf{supp}(\sigma_{k'}^*)$ with $k' > k$. So let us assume that $x_k > 0$. Then, for any $x < x_k$, Property A and the assumption of (strict) strategic complements imply that

$$v_{k+1}(x_k, x_{l_1}, \ldots, x_{l_k}, x_s) - v_{k+1}(x, x_{l_1}, \ldots, x_{l_k}, x_s) \geq v_k(x_k, x_{l_1}, \ldots, x_{l_k}) - v_k(x, x_{l_1}, \ldots, x_{l_k})$$

for any x_s, with the inequality being strict if $x_s > 0$. Then, averaging over all types, the fact that the degrees of any two neighbouring nodes are stochastically independent random variables together with the fact that there are some players with degree k who choose $x_k > 0$ implies that

$$U(x_k,\sigma^*, k + 1) - U(x, \sigma^*, k + 1) > U(x_k, \sigma^*, k) - U(x, \sigma^*, k).$$

On the other hand, note that from the choice of x_k,

$$U(x_k,\sigma^*, k) - U(x, \sigma^*, k) \geq 0$$

for all x. Combining the aforementioned considerations we conclude:

$$U(x_k,\sigma^*, k + 1) - U(x, \sigma^*, k + 1) > 0,$$

for all $x < x_k$. This in turn requires that if $x_{k+1} \in \mathbf{supp}(\sigma_{k+1}^*)$ then $x_{k+1} \geq x_k$, which of course implies that σ_{k+1}^* FOSD σ_k^*. Iterating the argument as needed, the desired conclusion follows, i.e., $\sigma_{k'}^*$ FOSD σ_k^* whenever $k' > k$. ■

Proof of Proposition 3. We present the proof for positive externalities. The proof for negative externalities is analogous and omitted. The claim is obviously true for a trivial equilibrium in which all players choose the action 0 with probability 1. So, let σ^* be a (non-trivial) equilibrium strategy. Suppose $x_k \in \mathbf{supp}(\sigma^*_k)$ and $x_{k+1} \in \mathbf{supp}(\sigma^*_{k+1})$. Property A implies that

$$v_{k+1}(x_k, x_{l_1}, \dots, x_{l_k}, 0) = v_k(x_k, x_{l_1}, \dots, x_{l_k}),$$

for all $x_{l_1}, \dots, x_{l_k}$. As the payoff structure satisfies positive externalities, it follows that for any $x > 0$,

$$v_{k+1}(x_k, x_{l_1}, \dots, x_{l_k}, x) \geq v_k(x_k, x_{l_1}, \dots, x_{l_k}).$$

We now have to consider two cases. First, assume positive neighbour affiliation and let σ^* be a monotone increasing equilibrium. Then, looking at expected utilities, we obtain that:

$$U(x_k, \sigma^*, k+1) \geq U(x_k, \sigma^*, k).$$

As σ^*_{k+1} is a best response in the network game being played and $x_{k+1} \in \mathbf{supp}(\sigma^*_{k+1})$,

$$U(x_{k+1}, \sigma^*, k+1) \geq U(x_k, \sigma^*, k+1)$$

and the result follows. Second, observe that the case of negative neighbour affiliation and monotone decreasing equilibrium strategy can be proven using analogous arguments. ■

Proof of Proposition 4. We know from Subsection 3.3 that network games of substitutes exhibit the degree substitutes property. Proposition 1 then tells us that there exists a symmetric equilibrium which is non-increasing in degree. Fix the strategy σ in one such equilibrium. Suppose that for degree $k > 0$ there is positive probability $\sigma(1|k)$ of choosing action 1. We prove that $\sigma(1|l) = 1$, for all $l < k$. Consider first degrees $l = k - 1 < k$. Then, letting the same notation $v_k(\cdot, \cdot)$ stand for the usual mixed extension of the original payoff function, the marginal return to action 1 can be written as follows:

$$\begin{aligned}
&U(1, \sigma, l) - U(0, \sigma, l) \\
&\quad = \sum_{(k_1,\dots,k_l)} P(k_1,\dots,k_l \mid l)[v_l(1,\sigma(k_1),\dots,\sigma(k_l)) - v_l(0,\sigma(k_1),\dots,\sigma(k_l))] \\
&\quad = \sum_{(k_1,\dots,k_{l+1})} P(k_1,\dots,k_l \mid l)[v_l(1,\sigma(k_1),\dots,\sigma(k_l),x_{l+1}=0) - v_k(0,\sigma(k_1),\dots,\sigma(k_l),x_{l+1}=0)] \\
&\quad \geq \sum_{(k_1,\dots,k_k)} P(k_1,\dots,k_k \mid k)[v_k(1,\sigma(k_1),\dots,\sigma(k_{k-1}),x_k=0) - v_k(0,\sigma(k_1),\dots,\sigma(k_{k-1}),x_k=0)]
\end{aligned}$$

$$> \sum_{(k_1,\ldots,k_k)} P(k_1,\ldots,kk \mid k)[v_k(1,\sigma(k_1),\ldots,\sigma(k_k)) - v_k(0,\sigma(k_1),\ldots,\sigma(k_k))]$$
$$= U(1,\ \sigma,\ k) - \mathrm{U}(0,\ \sigma,\ k) \geq 0,$$

where the second equality holds by Property A, the subsequent (weak) inequality holds because $\sigma(k - 1)$ FOSD $\sigma(k)$, P exhibits negative neighbour affiliation and strict strategic substitutes holds, and the second (strict) inequality holds due to strict strategic substitutes and $\sigma(1|k) > 0$. Finally, the last inequality simply reflects the hypothesis that σ constitutes an equilibrium. This argument can be repeated to establish that $\sigma(1|l) = 1$, for all $l < k$. Analogous arguments, with a simple switching of inequality signs, shows that if $\sigma(0|k) > 0$ then $\sigma(0|k') = 1$, for all $k' > k$.

The above argument establishes that every non-increasing symmetric equilibrium strategy σ is defined by a threshold t . To complete the proof, we next show that this threshold is unique. Thus, for the sake of contradiction, suppose that there are two distinct thresholds, t and t' with $t' < t$, which induce strategies σ and σ' respectively. If the equilibrium σ' is played, a player with degree $t' + 1$ (higher than the corresponding threshold t') strictly prefers action 0, i.e.

$$U(1,\ \sigma',\ t' + 1) - U(0,\ \sigma',\ t' + 1) < 0, \tag{A1}$$

while if equilibrium σ is played, a player with degree $t' + 1$ (no higher than the corresponding threshold t) weakly prefers action 1, i.e.

$$U(1,\ \sigma,\ t' + 1) - U(0,\ \sigma,\ t' + 1) \geq 0. \tag{A2}$$

We can then write:

$$0 \leq U(1,\ \sigma,\ t' + 1) - U(0,\ \sigma,\ t' + 1)$$
$$= \sum_{(k_1,\ldots,k_{t'+1})} P(k_1,\ldots,k_{t'+1} \mid t'+1)[v_{t'+1}(1,\sigma(k_1),\ldots,\sigma(k_{t'+1})) - v_{t'+1}(0,\sigma(k_1),\ldots,\sigma(k_{t'+1}))]$$
$$\leq \sum_{(k_1,\ldots,k_{t'+1})} P(k_1,\ldots,k_{t'+1} \mid t'+1)[v_{t'+1}(1,\sigma'(k_1),\ldots,\sigma'(k_{t'+1})) - v_{t'+1}(0,\sigma'(k_1),\ldots,\sigma'(k_{t'+1}))]$$
$$= U(1,\ \sigma',\ t' + 1) - U(0,\ \sigma',\ t' + 1) < 0$$

where the first and third inequalities are simply (A1) and (A2), while the middle inequality is a consequence of strategic substitutes and the hypothesis that $\sigma(1|k) \geq \sigma'(1|k)$, for all $k \in \{0, 1, \ldots, n - 1\}$. This yields the contradiction that completes the proof. ■

Proof of Proposition 5. Under the maintained hypotheses there exists a unique non-increasing symmetric equilibrium with a threshold property under both degree distributions. Suppose that this equilibrium σ' has threshold t' under P'. The assumptions that P dominates P' for all k and that players choose a non-increasing strategy imply that the equilibrium threshold under P cannot be lower than t'. To see this, suppose that in the non-increasing equilibrium under P, σ, the threshold $t < t'$. We now show that this yields a contradiction. In equilibrium σ' under P', for the threshold degree t' the expected payoffs from action 1 are higher than the expected payoffs from action 0. Thus, again identifying each $v_k(\cdot, \cdot)$ with the mixed extension of the corresponding payoff function, we can write:

$$\begin{aligned}
0 &\leq U(1, \sigma', t') - U(0, \sigma', t') \\
&= \sum_{(k_1,\dots,k_{t'})} P'(k_1,\dots,k_{t'} \,|\, t')[v_{t'}(1,\sigma'(k_1),\dots,\sigma'(k_{t'})) - v_{t'}(0,\sigma'(k_1),\dots,\sigma'(k_{t'}))] \\
&\leq \sum_{(k_1,\dots,k_{t'})} P(k_1,\dots,k_{t'} \,|\, t')[v_{t'}(1,\sigma'(k_1),\dots,\sigma'(k_{t'})) - v_{t'}(0,\sigma'(k_1),\dots,\sigma'(k_{t'}))] \\
&< \sum_{(k_1,\dots,k_{t'})} P(k_1,\dots,k_{t'} \,|\, t')[v_{t'}(1,\sigma(k_1),\dots,\sigma(k_{t'})) - v_{t'}(0,\sigma(k_1),\dots,\sigma(k_{t'}))] \\
&= U(1, \sigma, t') - U(0, \sigma, t'),
\end{aligned}$$

where the second inequality follows from the hypotheses that P dominates P', σ' is non-increasing and $v_{t'}(\cdot, \cdot)$ satisfies the strategic substitutes property, while the third inequality follows from the hypothesis that $t < t'$ and $v_{t'}(\cdot, \cdot)$ satisfies the strict strategic substitutes property. This however implies that for degree t' action 1 yields strictly higher expected payoffs than action 0 under equilibrium σ, a contradiction with $t < t'$. ■

Proof of Proposition 6. Suppose that $\tilde{F}(t') \leq \tilde{F}'(t'-1)$ but, contrary to what is claimed, $t < t'$. Then, under P, the probability that any of the neighbours chooses action 1 is bounded above by $\tilde{F}(t')$ and, therefore, by $\tilde{F}'(t'-1)$. Given the hypothesis that t' is the threshold under P', the assumption of strategic substitutes now generates a contradiction with the optimality of actions of degree t' in an equilibrium under P, and this completes the proof. ■

Proof of Proposition 10. Let us consider first the case of strategic complements and denote by $\sum^m$ the set of monotone strategies. The proof is based on the following two claims:

Claim 1. For any player i, if all other players $j \neq i$ use a common strategy $\sigma \in \sum^m$ there is always a strategy $\sigma_i \in \sum^m$ that is a best response to it.

Claim 2. A symmetric equilibrium exists in the strategic form game where players' strategies are taken from $\sum^m$.

To establish Claim 1, consider a player i and let $t_i, t_i' \in \mathcal{T}$ such that $t_i' \succeq t'$, where $\succeq$ is the partial order applicable to the case of strategic complements (see Section 6). For any $\sigma \in \sum^m$ chosen by every $j \neq i$, let $BR(\sigma, t_i)$ be the set of best-response strategies of player i to σ when his type is t_i. Let us assume that $\sigma(t_j) \neq 0$

for some $t_j \in \mathcal{T}$. (Otherwise, the desired conclusion follows even more directly, since the best-response correspondence is unaffected by being connected to a player whose strategy chooses action 0 uniformly.) By definition, for every $x_{t_i} \in$ BR(σ, t_i), we must have that

$$\forall x \in X, \quad U(x_{t_i}, \sigma, t_i) - U(x, \sigma, t_i) \geq 0.$$

Then, since $t_i' \succeq t_i$, the assumption of (strict) strategic complements implies that

$$\forall x \in x_{t_i}, \quad U(x_{t_i}, \sigma, t_i') - U(x, \sigma, t_i') > 0. \tag{A3}$$

This follows from a two-fold observation:

(i) From Property A, if $t_i = (k, \ell_1, \ell_2, ..., \ell_k)$ and $t_i' = (k', \ell_1', \ell_2', ..., \ell_{k'}')$ and $t_i' \succeq t_i$ we can think of t_i involving k' neighbours with all neighbours indexed from $k + 1$ to k' (if any) choosing the action 0;

(ii) From strict strategic complements, since $\ell_u' \geq \ell_u$ the probability distribution over actions corresponding to each of his neighbours under t_i, $u = 1, 2, ..., k$, is dominated in the FOSD sense by the corresponding neighbour under t_i'. This follows from the fact the beliefs applying separately to each of the ℓ_u and ℓ_u' second-neighbours under consideration in each case are identical and stochastically independent.

Let us now make use of (A3) in the case where x_{t_i} is the highest best response by type t_i to σ. Then, it follows that any $x_{t_i'} \in BR(\sigma, t_i')$ must satisfy:

$$x_{t_i'} \geq \sup\{ x_{t_i} : x_{t_i} \in BR(\sigma, t_i)\},$$

which establishes Claim 1.

To prove Claim 2, we can simply invoke that, for any given $x \in X^k$, the function $v_k(\cdot, x)$ in own action either has a discrete domain or is concave, combined with the fact that the set of monotone strategies is compact and convex. To see the latter point, note that the monotonicity of a strategy σ is characterized by the condition:

$$\forall t_i, t_i' \in \mathcal{T}, \quad t_i' \succeq t_i \Rightarrow \sigma(t_i') \text{ FOSD } \sigma(t_i). \tag{A4}$$

Clearly, if two different strategies σ and σ' satisfy (A4), then any convex combination $\hat{\sigma} = \lambda\sigma + (1 - \lambda)\sigma'$ also satisfies it.

Finally, to prove the result for the case of strategic substitutes, note that the above line of arguments can be applied unchanged, with the suitable adaptation of the partial order used to define monotonicity. In this second case, as explained in Section 6, we say that $t \succeq t'$ if and only if $k \geq k'$ and $\ell_u \leq \ell_u'$ for all $u = 1, 2, ..., k'$. ■

References[23]

Ballester, C., A. Calvó-Armengol and Y. Zenou (2006), 'Who's who in networks. Wanted: The key player', *Econometrica* **74**(5), 1403–1417.

Barabási, A.L. and R. Albert (1999), 'Emergence of scaling in random networks', *Science* **286**, 509–512.

Bramoullé, Y. and R. Kranton (2007), 'Public goods in networks', *Journal of Economic Theory* **135**(1), 478–494.

Burt, R.S. (1994), *Structural Holes*, New York: Academic Press.

Calvó-Armengol, A. and M.O. Jackson (2009), 'Like father, like son: Labor market networks and social mobility', *American Economics Journal: Microeconomics* **1**(1), 124–150.

Carlsson, H. and E. Van Damme (1993), 'Global games and equilibrium selection', *Econometrica* **61**(5), 989–1018.

Chwe, M.S.Y. (2000), 'Communication and coordination in social networks', *Review of Economic Studies* **65**, 1–16.

Coleman, J. (1966), *Medical Innovation: A Diffusion Study*, 2nd edn, New York: Bobbs-Merrill.

Conley, T. and C. Udry (2010), 'Learning about a new technology: Pineapple in Ghana', *American Economic Review* **100**(1), 35–69.

Ellison, G. (1993), 'Learning, local interaction, and coordination', *Econometrica* **61**, 1047–1071.

Erdös, P. and A. Rényi (1960), 'On the evolution of random graphs', *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17–61.

Esary, J.D., F. Proschan, and D.W. Walkup (1967), 'Association of random variables, with applications', *Annals of Mathematical Statistics* **38**(5), 1466–1474.

Feick, L.F. and L.L. Price (1987), 'The market maven: A diffuser of marketplace information', *Journal of Marketing* **51**(1), 83–97.

Foster, A.D. and M.R. Rosenzweig (1995), 'Learning by doing and learning from others: Human capital and technical change in agriculture', *Journal of Political Economy* **103**(6), 1176–1209.

Galeotti, A. (2008), 'Talking, searching and pricing', Mimeo, University of Essex.

Galeotti, A., S. Goyal, M.O. Jackson, F. Vega-Redondo and L. Yariv (2006), 'Network games', Mimeo, University of Essex and Caltech.

Galeotti, A. and F. Vega-Redondo (2006), 'Complex networks and local externalities: A strategic approach', Mimeo, Universities of Essex and Alicante.

Glaeser, E., B. Sacredote and J. Scheinkman (1996), 'Crime and social interactions', *Quarterly Journal of Economics* **111**, 507–548.

Glaeser, E.L. and J. Scheinkman (2003), 'Non-market interactions', in M. Dewatripont, L.P. Hansen and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Cambridge, UK: Cambridge University Press.

Goyal, S. (2007), *Connections: An Introduction to the Economics of Networks,* Princeton: Princeton University Press.

Goyal, S. and J.L. Moraga-Gonzalez (2001), 'R&D networks', *Rand Journal of Economics* **32**(4), 686–707.

Granovetter, M. (1994), *Getting a Job: A Study of Contacts and Careers*, Evanston: Northwestern University Press.

Hirshleifer, J. (1983), 'From weakest-link to best-shot: The voluntary provision of public goods', *Public Choice* **41**(3), 371–386.

[23] [Editor's note] Conley and Udry (2010), which was originally cited as forthcoming, has been updated.

Jackson, M.O. (2008), *Social and Economic Networks*, Princeton: Princeton University Press.

Jackson, M.O. and B. Rogers (2007), 'Meeting strangers and friends of friends: How random are socially generated networks?', *American Economic Review* **97**(3), 890–915.

Jackson, M.O. and L. Yariv, (2005), 'Diffusion on social networks'*, Économie Publique* **16**, 3–16.

Jackson, M.O. and L. Yariv (2007), 'Diffusion of behavior and equilibrium properties in network games', *American Economic Review* (Papers and Proceedings) **97**(2), 92–98.

Jackson, M.O. and L. Yariv (2008), 'Diffusion, strategic interaction, and social structure', in J. Benhabib, A. Bisin and M.O. Jackson (eds.), *Handbook of Social Economics*, Elsevier.

Kakade, S., M. Kearns, J. Langford and L. Ortiz (2003), *Correlated Equilibria in Graphical Games,* New York: ACM Conference on Electronic Commerce.

Katz, E. and P.F. Lazarsfeld (1955), *Personal Influence: The Part Played by People in the Flow of Mass Communication*, Glencoe: Free Press.

Kearns, M., M. Littman and S. Singh (2001), 'Graphical models for game theory', in J.S. Breese and D. Koller (eds.), *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, San Francisco: Morgan Kaufmann University of Washington, pp. 253–260.

Kets, W. (2007), 'Beliefs in network games', CentER Discussion Paper 2007-46.

Kumbasar, E., A.K. Romney and W. Batchelder (1994), 'Systematic biases in social perception', *American Journal of Sociology* **100**, 477–505.

Lehmann, E.L. (1966), 'Some concepts of dependence', *The Annals of Mathematical Statistics* **37**, 1137–1153.

Milgrom, P. and C. Shannon (1994), 'Monotone comparative statics', *Econometrica* **62**, 157–180.

Milgrom, P. and R.J. Weber (1982), 'A theory of auctions and competitive bidding', *Econometrica* **50**, 1089–1122.

Morris, S. (2000), 'Contagion', *The Review of Economic Studies*, **67**(1), 57–78.

Morris, S. and H. Shin (2003), 'Global games: Theory and applications', in M. Dewatripont, L. Hansen and S. Turnovsky (eds.), *Advances in Economics and Econometrics*, Proceedings of the Eighth World Congress of the Econometric Society, Cambridge: Cambridge University Press, pp. 56–114.

Newman, M.E.J. (2003), 'The structure and function of complex networks', *SIAM Review* **45**, 167–256.

Sundararajan, A. (2006), 'Local network effects and network structure', *The B.E. Press Journal of Theoretical Economics* **6**(1) (Contributions).

Topa, G. (2001), 'Social interactions, local spillovers and unemployment', *Review of Economic Studies* **68**(2), 261–295.

Van Zandt, T. and X. Vives (2007), 'Monotone equilibria in Bayesian games of strategic complementarities', *Journal of Economic Theory* **134**, 339–360.

Vazquez, A. (2003), 'Growing network with local rules: preferential attachment clustering hierarchy, and degree correlations', *Physical Review E* **67**, 056104.

Vega-Redondo, F. (2007), *Complex Social Networks, Econometric Society Monograph,* Cambridge, UK: Cambridge University Press.

Weinstein, D. and M. Yildiz (2007), 'A structure theorem for rationalizability with application to robust predictions of refinements', *Econometrica* **75**(2), 365–400.

Network Organizations

Fernando Vega-Redondo

Paper presented at the CTN Workshop 2008. Reprinted from *Journal of Complex Networks,* 2013, **1**(1), June 2013, 72–82, with permission from Oxford Journals.

It is common to define a network organization as one that is fast and flexible in adapting to changes in the underlying environment. But besides the short-run advantages of adaptability, fast changes in the structure of the organization can also be detrimental in the longer run. This is due to the fact that agents need some stability in the organizational structure in order to channel appropriately (and thus speed up) search.

I discuss that trade-off between adaptability and structural stability in a context where not only the environment is continuously changing over time but the organization is also adjusting to those changes. The main conclusion obtained is that, as the environmental volatility increases, the optimal functioning mode of the organization sharply switches from being totally flexible to being completely rigid, i.e. no intermediate configurations are essentially ever optimal. This has stark positive and normative implications on the dichotomy of stability versus change that is at the center of recent organization literature.

1. Introduction

A ‘network organization’ is usually conceived as an organization that is quick and flexible in adapting to changes in its environment. But changes in the structure of the organization can also be detrimental in the medium run, since it is partly the

FV-R grateful to Bruce Kogut and participants in seminars at INSEAD, the University of Essex, and the 2008 CTN workshop at Venice for helpful comments. FV-R also wants to thank Roman Chuhay for conducting the numerical simulations used to support the theoretical analysis of the model.

Funding: FV-R acknowledges financial support from the Spanish Ministry of Education under grant SEJ2007-62656.

knowledge of the organization's structure that mediates (and thus speeds up) search. Here I discuss the tension between these two considerations. That is, I study the trade-off between adaptability and structural stability in a (network) organization that confronts a changing environment.

The model proposed to study this trade-off is particularly simple and stylized. The organization consists of an underlying backbone structure (a one-dimensional lattice network) that remains fixed, combined with a limited number of links that can be 'rewired' over time (for simplicity, just one per agent). In the background, there is an environment that changes over time – a phenomenon that we call volatility. More specifically, it is assumed that every node/agent has a target node it has to reach, whose identity independently changes at a rate p. The dilemma faced by a node whose target has been reassigned is the following: should I redraw my flexible link to the new target? If this is done, direct access to that node (possibly the target as well in the immediate future) is secured. But, on the other hand, under the assumption that freshly rewired links take some time to become widely available to the organization at large, such an adaptation also imposes a negative externality on others. Namely, it removes from the immediate operational structure of the organization some links that can be particularly valuable for overall search.

So, in a nutshell, the problem we pose can be formulated as follows. What is the optimal speed at which the organization should adapt to the changing environment? To cast the question sharply, the adaptability of the organization is supposed embodied by a single parameter q, the probability with which a node will redraw its flexible link to a new target. In this setup, the answer delivered by the paper is a drastic one: depending on whether the value of p (volatility) is high or low, the optimal q (adaptability) should essentially be, respectively, either zero or one. Thus, in this sense, one finds that an optimal organization is typically either totally rigid or totally flexible, and essentially never in between. As we shall explain, this has an interesting bearing on the dichotomy of stability versus change that has been highlighted by recent organization literature.

The rest of the paper is organized as follows. Next, Section 2 provides a brief discussion of related literature. Then, Section 3 presents the model. The analysis is undertaken in Section 4, while Section 5 concludes.

2. Related Literature

There are three distinct branches of literature quite related to our present concerns: (a) the economic theory of organizations, (b) models of search in complex networks, (c) the transactive-memory theory of organizations. I briefly discuss each of them in turn.

(a) The economic theory of organizations has produced a large body of research whose focus has been both on incentive issues and/or the way in which organizations can effectively handle decentralized information. In the latter vein, the work of Radner (1993) was a seminal contribution that (abstracting from incentive considerations) first modelled explicitly the organization as a network of informal flows. Other subsequent researchers have followed his lead (see e.g. Bolton and Dewatripoint (1994), van Zandt (1999), and Garicano (2000)), all aiming at characterizing the optimal network structure that, under varying conditions and in different senses, best pools the information disseminated throughout the organization. Even more in line with our approach, a recent interesting paper that stresses the issue of organization adaptability in the face on environmental change is Dessein and Santos (2006). Their focus, however, is on the tradeoff between coordination and specialization when individuals have only local information on the environment and their communication is impaired by noise.

(b) In recent years, and partly motivated by the rise to prominence of internet, there has been a surge of interest on the problem of how to conduct effective search in large and complex social networks. Building upon the early experimental work of Milgram (1967) on ‘small worlds’ and the subsequent theoretical developments of Watts and Strogatz (1998), a key issue in this respect is that of searchability. More specifically, the question is *how to find* short paths joining the nodes of large networks that (as indeed happens in the real world) involve a significant random component. Kleinberg (2000) provided key insights on the problem, formulating it as one of an algorithmic nature. The path opened by this seminal contribution has then been pursued by several authors – see e.g. Guimerà et al. (2002) or Dodds et al. (2003) – to address issues of organization design. Specifically, they pose the problem of how to design the social network underlying the operation of large organizations so as to optimize their search-related performance. In contrast with our approach, however, the underlying network is taken as fixed once and for all, so that the notion of adaptability does not pertain to the organizational structure governing informational flows.

(c) Finally, I discuss the so-called transactive-memory theory of organizations. This theory originates in the work of Wegner (1986). He stressed the importance of the process by which, as new information arrives to an organization, it is first allocated to individuals, then registered in the ‘organizational directory’, and later retrieved in the most efficient manner. This three-fold mechanism is what has been called the organization’s transactive memory system. A large body of theoretical and empirical literature has followed suit (see e.g. Moreland and Argote (2003) for a survey). In much of it, researchers have emphasized the importance of informal (and thus flexible) links in the successful implementation of an organization’s transactive memory.

A good case in point is provided by the empirical work of Hansen (2003).[1] He

[1] See also Hansen (1999) and Schulz (2003).

studied 120 product development projects in a large electronics company, where each project was separately undertaken by one of the 41 business units of the firm. Hansen started by constructing a knowledge network, on the basis of the informal contacts identified among the members of the different units. Then, much in line with the key assumptions of our model, he found that the overall performance of each unit (specifically, the fraction of projects completed and the speed of their completion) was highly dependent on the existence of short network paths to other units possessing relevant knowledge. Indirect connections, in other words, were crucial for good results, but their value was found to decay significantly with distance.[2] This, indeed, is consistent with the central measure of performance contemplated in our model, which in turn underlies the problem of network design addressed by our theoretical analysis.

3. The Model

Consider a large set of nodes $N = \{1, 2, ..., n\}$ arranged along an organizational backbone, which is assumed to be one-dimensional and without boundaries, i.e. a ring.[3] Each node i is connected to both of its direct neighbors in the backbone, $i - 1$ and $i + 1$ (where the index here is interpreted as 'modulo n'). These links are conceived as formal and rigid ones, possibly reflecting the *formal chart* of the organization. In addition to such formal links, every i is also connected to some $\alpha(i)$ through an informal link, which may well be 'long-range' (i.e. far away from i on the underlying backbone). Such *long-range links* are flexible so they can be adjusted over time, as determined by the plasticity/adaptability of the organization (see below for the dynamic formulation). For the moment, we may simply assume that each $\alpha(i)$ has been randomly selected from $N\backslash\{i\}$ with uniform probability. The resulting (undirected) network – consisting of the backbone plus the long-range links – will be denoted by Γ.

Let us further postulate that each node $i \in N$ has a *target* $\tau(i) \in N\backslash\{i\}$, whom i has to reach in order to address a specific demand or tackle a particular problem. Again, let us momentarily assume that $\tau(i)$ has been randomly selected from $N\backslash\{i\}$ with uniform probability. Then, given the prevailing network Γ and the array of targets $\tau = [\tau(i)]_{i=1}^{n}$ the performance of the organization is tailored to the

[2] These considerations were most decisive for the transfer of knowledge that could be largely codified. Instead, for hardly codifiable knowledge, direct and close contact between the source and the target played a primary role.

[3] Conceivably, one could consider other network architectures – e.g. hierarchic or tree-like – to model the backbone of the organization. This, however, would complicate the formal analysis of the setup, which in our case heavily relies on the theory that has been developed for the so-called small-world networks, i.e. networks defined on a lattice (in terms of the corresponding distance), to which a few lattice-independent links are added to establish some extent of 'global connectivity'. It seems intuitive, however, that none of the essential features and insights of the model depend on the details posited for the fixed backbone of the organization.

average path length (along the network) between every node and its target. More specifically, its aim will be to have this magnitude be as low as possible, so that the objective function that the organization will aim to maximize is given by

$$\rho = -\langle d(i, \tau(i)) \mid \Gamma \rangle_{i \in N}.$$

This is motivated by the idea that the network distance separating an agent from a valuable partner (e.g. one that helps undertake current tasks) should have an important bearing on the speed and success of job completion. As explained above – recall the Introduction – this idea is not only intuitive but also enjoys some significant empirical support.

But, as advanced, the focus of the paper is on the tension between adaptability and structure in a *dynamic context* where the environment changes over time. So let us introduce time into the model, indexing it by $t = 0, 1, 2, \ldots$ and dating the prevailing states $\omega_t \equiv [\tau_t(i), \Gamma_t\ ]$ accordingly. Suppose that the initial state $\omega_0 \equiv [\tau_0(i), \Gamma_0]$ is constructed randomly, as explained above – i.e. both the target and the long-range neighbor of each node are selected in a stochastically independent and uniform manner among all the other nodes. Then, as time proceeds, the law of motion that governs the change from ω_{t-1} to ω_t for every $t \geq 1$ consists of two separate components: target revision and update of the long-range neighbor. For simplicity, we assume that each of these components is implemented sequentially in the following two consecutive stages:

1. *Target revision*: Independently (i.e. 'simultaneously') for each node i, its previous target $\tau_{t-1}(i)$ is redrawn afresh with probability $p \in [0,1]$. (Thus, with probability $(1-p)$, we have $\tau_t(i) = \tau_{t-1}(i)$.) When a new target for i is redrawn, each $j \in N\backslash\{i\}$ is selected as the new target with uniform probability. (So, in principle, any given node can act as the target for several other nodes.)
2. *Neighbor update*: Independently for each node i, its previous long-range link to agent $\alpha_{t-1}(i)$ is rewired with probability q to the current target $\tau_t(i)$. (Thus, with probability no lower than q at every t, the long-range link of i connects to its current target, i.e. $\alpha_t(i) = \tau_t(i)$.)

The first component of the law of motion, *target revision*, embodies the idea of volatility: over time, the environment evolves and the needs/tasks/objectives of individual nodes are affected by it. The parameter $p \in [0,1]$ modulates the rate at which the environment changes, thus leading to pressure for some adjustment to take place.

On the other hand, the second component, *neighbor update*, specifies how and when, in response to the aforementioned adjustment pressure, actual changes in the network structure unfold as agents change their long-range neighbors. The parameter $q \in [0,1]$ is a measure of organizational plasticity. It can be conceived as an attribute of organizations – say, a part their 'culture' – and will generally differ across them.[4] Sometimes, it may also be regarded as an outcome of design, at least partially. This, for example, is what is implicitly suggested when managers or consultants speak of reshaping the culture of a firm in order to improve its performance.

Finally, we turn to the issue of how to *measure performance* in the dynamic setup. As explained and motivated above, organizational performance is associated to the average distance between nodes and their targets. But, in the present dynamic context, we want to add a key twist to it. Specifically, we posit that, in computing node-target distances, only the links in $\Gamma_{t-1} \cap \Gamma_t$ can be used. That is, only the informal links that have remained in place for at least one period are assumed to form part of the *operational communication structure* of the organization. (At the beginning of the process, we posit that $\Gamma_0 = \Gamma_{-1}$, so that all initial links form part of the organizational structure.)

Several (complementary) justifications can be given to this assumption. One is that some 'socialization' time (here, just one period) is required for a fresh link to be formed and become effective. An alternative motivation is that it also takes time for a new link between two individuals to be known (and thus become usable) by the rest of the organization in accessing their targets. Formally, the implication of this assumption is that organizational performance ρ_t at each t is to be measured as follows:

$$\rho_t = -\langle d(i,\ \tau_t(i)) \mid \Gamma_{t-1} \cap \Gamma_t \rangle_{i \in N}, \tag{1}$$

where the conditioning included in the average $\langle \cdot \rangle$ indicates that, in computing the distances $d(\cdot)$, only links in $\Gamma_{t-1} \cap \Gamma_t$ can be used.

Clearly, it is the delay in the effectiveness of new links contemplated in (1) that introduces the trade-off between adaptability and structure that is at the heart of our model. If no such delay prevailed, instantaneous adaptability to any change in the environment (i.e. a value of $q = 1$) would obviously maximize organizational performance. Instead, when some time must elapse between the establishment of a new link and its effective use by the organization, an interesting tension arises. On the one hand, there is the *immediate* benefit to the individual adjusting node from having a link (and hence direct access) to its new target. And on the other hand, after any such adjustment by an individual node has taken place, the organization as a whole must face the cost imposed on other nodes by the *temporary* reduction of their communication structure.[5] In general, as a result of these conflicting

[4] This is stressed in the influential work of Schein (2002, 2004), who conceives culture as the background for change in any organization. In fact, somewhat in line with the role of q in our model, he suggests that the 'shared assumptions and beliefs about the stability of human relationships' is a key cultural dimension that differentiates organizations.

[5] The same tradeoff and qualitative implications would arise if, rather than one period, the (finite) delay involved in a new link becoming part of the communication structure of the organization were longer.

considerations, it may be optimal for the organization to limit its adaptability to the recurrent changes in the environment (i.e. display a value of $q < 1$).

4. Analysis

In a nutshell, our main objective will be to shed light on the interplay between the *plasticity* of the organization (as given by q) and the *volatility* of its environment (as captured by p). To fix ideas, a useful way to grasp this relationship is to consider an optimal-design problem in which p is the exogenous parameter and q is the decision/design variable. Naturally, this problem must be formulated in a long intertemporal framework, where volatility and adjustment have a full chance to unfold. Thus, let us take a truly long-run perspective and identify the overall performance of the organization with

$$\rho \equiv \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \rho_t .$$

Since the underlying stochastic process is ergodic, ρ is independent of initial conditions and can be conceived as a function of p and q alone. Let us write $\rho(p, q)$ to reflect such dependence. Then, our theoretical concerns are addressed by the following optimization problem: given any $p \in [0,1]$, find $q^*(p)$ such that

$$q^*(p) \in \arg\max_{q\in[0,1]} \rho(p,q). \tag{2}$$

In essence, this optimization problem reflects a trade-off between two opposing objectives:

1. *adaptability* – swiftness in responding to a mismatch between links and targets;
2. *structure* – preservation of the network connectivity (specifically, the long-range links) required to conduct search effectively.

To gain an analytical understanding of the essential implications resulting from this trade-off, we study the problem through an idealization of our framework in which the dynamics of the system is identified with its expected motion. As customary, we call such an idealization the Mean-Field Model (MFM). Given the stochastic independence displayed by the forces impinging on each node (both concerning volatility and link rewiring), it is natural to conjecture that the MFM should capture the essential behavior displayed by large finite systems. Indeed, this will be confirmed by numerical simulations, whose performance are found to match with great accuracy the theoretical predictions.

The MFM is defined by a dynamical system formulated on a population-wide (anonymous) description of the evolving situation. A sufficient specification (or state) of this system is given by the fraction of nodes that are currently on target – i.e. are connected to their respective target through their long-range links. Let $\mu(t)$ stand for the fraction of such nodes prevailing at some t. Then, its expected law of motion is given by the following simple difference equation:

$$\mu(t+1) = (1-p)\mu(t) + q[\mu(t)p + (1-\mu(t))].$$

It is straightforward to see that the system globally converges to a unique positive fraction of nodes on target given by

$$\mu^* = \frac{q}{p+q(1-p)} < 1. \tag{3}$$

This implies that, in the long run, the MFM predicts that the total number of long-range links that are fully operational (i.e. have been in place for at least one period) is given by:

$$\begin{aligned}\lambda^* &\equiv [(1-p)\mu^* + (1-q)(1-(1-p)\mu^*)]n \\ &= \frac{p(1-q)+q(1-p)}{p+q(1-p)}n.\end{aligned} \tag{4}$$

Our next step is to compute the long-run average distance between a node and its target, when a direct link does not already exist between them. To this end, we note that the 'operational network' prevailing at t (i.e. $\Gamma_{t-1} \cap \Gamma_t$) can be conceived as a small-world network of the sort studied by Newman et al. (2000), itself a variation of the original setup proposed by Watts and Strogatz (1998). Very succinctly, such a network is constructed as follows:

(a) One starts with a large set of nodes, arranged linearly along a ring. Each of them is taken to be connected to the two nodes adjacent to it along the ring.[6]

(b) Then, independently across every node, each of them is given an additional 'short-cut' with some (small) probability φ. This short-cut is a link that connects the node in question to some randomly selected node in the whole set.

Formally, the number of short-cuts in the small-world network can be identified with the number λ of operational long-range links in our setup. This then allows one to rely directly on the expression derived by Newman et al. (2000) to approximate the average network distance in their small-world setup. In terms

[6] In the general original formulation, the nodes can be directly connected to all those that lie within a certain number of steps away in the ring. This generalization is irrelevant for our purposes.

of our present notation, they found it to be proportional to a function $F(\lambda)$ given by

$$F(\lambda) = \frac{2}{\sqrt{\lambda^2 + 2\lambda}} \tanh^{-1}\left(\frac{\lambda}{\sqrt{\lambda^2 + 2\lambda}} \right).$$

The function above, of course, only applies to the nodes that are not on target. Since the fraction of these in the long run is $[1 - (1 - p)\mu^*]$, the objective function ψ to be maximized in our case is[7]

$$\psi(p, q) = - [1 - (1 - p)\mu^*(p, q)]F(\lambda^*(p, q)). \tag{5}$$

where $\mu^*(p, q)$ and $\lambda^*(p, q)$ stand for the long-run values for μ and λ respectively given by (3) and (4), the notation highlighting that they both depend on the parameters of the model, p and q.

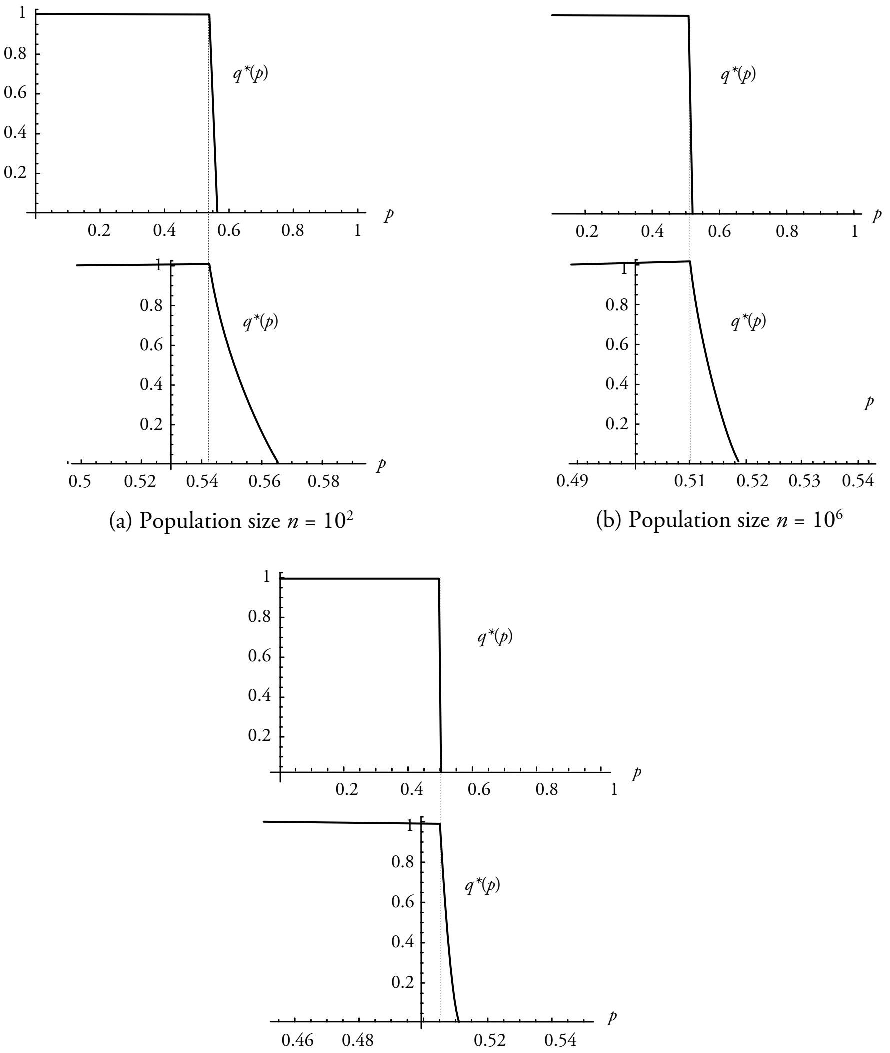
Combining the previous considerations, the optimization problem faced by the organization can be formulated as follows. Given any $p \in [0,1]$ (the environmental volatility), find the value $q^*(p)$ (optimal plasticity) that solves

$$\max_{q \in [0,1]} \psi(p, q) \,. \tag{6}$$

Once the full dependence on p and q is taken into account, the function $\psi(p, q)$ becomes rather involved, which makes it hard to characterize analytically the solution of the above optimization problem. I choose, therefore, to rely on numerical methods (as implemented e.g. by standard software packages) to identify the optimal plasticity q^* that solves (6), as a function of the volatility rate p. Figure 1 describes the induced mapping $q^*(p)$ for different values of population size and show that it is qualitatively the same across a wide range in orders of magnitude.

The results depicted in Figures 1(a-c) provide a stark picture of the way in which the tension between structure and adaptability is resolved in a network organization that is suitably described by our model. It shows that, except for a very narrow transition range, the optimal level of organizational plasticity is either full (i.e. $q^*(p) = 1$) or completely absent (i.e. $q^*(p) = 0$). Thus, if we focus on the rate at which the organization effectively changes its network structure, the conclusions can be described as follows. For low levels of volatility, the rate of change matches that of the environment since the plasticity of the organization is maximal. Thus, as the environment gets more volatile, the organization undergoes

[7] In line with the model proposed by Newman et al. (2000), we gain some notational simplicity by normalizing the distance between direct neighbors in the network to zero. This implies that nodes that are directly connected to their targets can be ignored in the performance measure as they lead to no cost or delay in tackling the corresponding problems.



(a) Population size $n = 10^2$

(b) Population size $n = 10^6$

(c) Population size $n = 10^{10}$

Figure 1. Optimal plasticity $q^*(p)$ as a function of volatility p for various population sizes, $n = 10^2, 10^6, 10^{10}$. The function is shown both for the whole domain $p \in [0,1]$ as well as for a scaled version that is 'zoomed in' on the region where the transition from high to low optimal values takes place.

a fully parallel (linear) increase in network adjustment. This state of affairs, however, ends abruptly at levels of volatility well below complete target turnover. For, at a value of p sizably below 1, the optimal plasticity of the organization falls steeply to zero. There is, therefore, a wide range for p in which the best performance is achieved by freezing the network of the organization at its original random configuration.

The analytical solutions derived from the MFM closely match the behavior observed in numerical simulations of the model, even if the population is relatively small. By way of illustration, Figure 2 shows simulation results for $n = 100$, which can be compared with the theoretical prediction depicted in Figure 1(a) for the same population size. In both cases we observe that, within a relatively narrow interval for p that lies above ½, there is a sharp transition across extreme degrees of organizational plasticity (i.e. probabilities $q \in \{0,1\}$). In the simulations the transition is fully completed along the interval [0.55,0.62], while the theory displays a narrower transition range contained in [0.54,0.57].

Our conclusions shed light on points made, in diverse forms, by the recent organization literature. For example, Schein (2002, 2004) argues that stability and change are 'two sides of the same coin', and that both are part of any successful adaptation to an environment in perpetual flux. Moreland and Argote (2003), on the other hand, elaborate on this idea by emphasizing that too much flexibility may deteriorate the so-called 'intellectual capital' of the organization (i.e. the knowledge available to an organization through its workers). This capital is

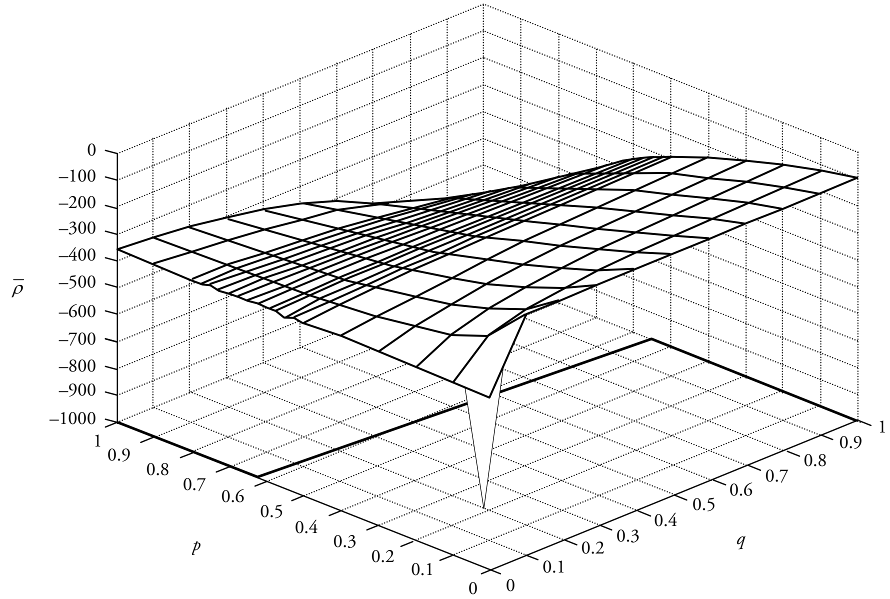


Figure 2. The upper surface depicts the average performance $\bar{\rho} \equiv (1/T)\sum_{t=1}^{T} \rho_t$ *over* $T = 20000$ *rounds in a context consisting of* $n = 100$ *agents where each* ρ_t *is computed as in (1). The lower line on the* $p - q$ *plane represents the optimal plasticity* q *for which* $\bar{\rho}$ *is maximized at each of the volatility rates p considered (a grid with step value* $\Delta = 0.025$*). The transition from a situation with full plasticity (*$q = 1$*) to another with none at all (*$q = 0$*) occurs as volatility (the probability p) grows from* $p = 0.550$ *to* $p = 0.625$*. As explained in the text, this interval is similar to that predicted by the theory (cf. Figure 1a) but somewhat wider.*

accessed by the organization's transactive-memory system – recall Section 2 – whose operation is crucially facilitated by 'a shared awareness among workers of who knows what (...)'.

The model may be regarded as having both descriptive and normative implications. On the descriptive side, one of its predictions is that, to the extent that organizations can be taken to operate efficiently, the most rigid ones should be those operating in the most volatile environments. This, however, raises normative issues as well, bearing on the likely conflict between individual incentives to adjust and the possibly detrimental effects of such an adjustment on the overall performance of the organization.

Our present approach does not take individual payoffs into account, and thus precludes a rigorous study of such normative questions. Any extended model that would do so, however, should probably posit that individual incentives to adjust long-range links are directly related to the current distance between node and target. Then, if one were to abstract from any adjustment costs, maximum plasticity would always be optimal from a purely individual perspective. But, as our analysis underscores, this may be suboptimal for the organization as a whole if the volatility of the environment is relatively high. In essence, the problem at stake is a classical one of externalities – in this case, externalities of individual adjustment on the search effort by others. And, as usual, what the problem may then require is a suitable kind of intervention that, by impinging on individuals ability or/and payoffs to adjust, leads to a socially optimal outcome. To formulate and analyze this 'implementation problem' in any detail is outside the scope of the present paper.

5. Summary and Future Research

The paper studies a model of a network organization that lives in a volatile environment and must therefore face the trade-off between the adaptability to changing circumstances and the preservation of an operational network structure. Our analysis yielded rather clear-cut conclusions. Specifically, we found that the positive effects of adaptability fully dominate for low levels of volatility but are sharply and completely offset beyond some intermediate threshold. This raises positive and normative issues on the 'dynamic design' of organizations, which are left for subsequent research.

Additional issues to be explored in the future concern the sensitivity of these conclusions to some of the simplifying features of the approach. Since the theoretical framework is so stylized, many extensions could be explored. By way of illustration, consider the assumption that the rewiring of a long range link occurs with the same probability, independently of the distance to the target that is closed by the adjustment. In the same spirit of the model, it would be natural to postulate instead that rewiring occurs (say, again with some probability q) only if that distance exceeds a certain threshold. Obviously, a suitable choice for this threshold could just improve the overall performance of the organization. But a trade-off akin to that of the original model would still arise and, therefore, it would be interesting to know whether similar conclusions continue to hold, at least qualitatively.

In my view, however, one of the most interesting variations of the model to be considered would affect the postulated backbone of the organization. The present model has assumed that this backbone is a regular boundariless lattice (i.e. a 'ring'). Often, however, the formal and stable network of an organization is best conceived as displaying a less symmetric form. A natural alternative is given by a hierarchical tree structure, where each individual – except for the single apex – is connected to one 'supervisor'. Such a hierarchy is descriptive of many of the real-world structures observed in organizations, and probably this is partly due to the advantages it allows in the routing and processing of information (cf. Radner, 2003). Recently, however, it has been argued (see e.g. Dodds et al., 2003) that the addition of long-range links connecting distant parts of an underlying hierarchic structure can greatly improve its overall performance. Indeed, this is supported by a large body of empirical research which finds that '(...) much of the real work in any company gets done through an informal organization, with complex networks of relationships that cross functions and divisions'. (Cf. Krackhardt and Hanson ,1993)

The models that have been proposed in the theoretical literature to understand the aforementioned considerations, however, have been mostly static. They conceive the organization network as fixed, even if it consists of a complex blend of hierarchic and transversal links. To enrich that approach with a genuinely dynamic model of the organization appears to be an interesting development, which could be carried out along the lines suggested in this paper.

References

Bolton, P. and M. Dewatripont (1994), 'The Firm as a Communication Network', *Quarterly Journal of Economics* **109**, 809–839.

Dessein, W. and T. Santos (2006), 'Adaptive organizations', *Journal of Political Economy* **14**, 956–995.

Dodds, P.S., D.J. Watts and C.F. Sabel (2003), 'Information exchange and the robustness of organizational networks', *Proceedings of the National Academy of Sciences* **100**, 12516–12521.

Garicano, L. (2000), 'Hierarchies and the Organization of Knowledge in Production', *Journal of Political Economy* **108**, 874–904.

Guimerà, R., A. Díaz-Guilera, F. Vega-Redondo, A. Cabrales and A. Arenas (2002), 'Optimal network topologies for local search with congestion', *Physical Review Letters* **89**, 248701.

Hansen, M.T. (1999), 'The search-transfer problem: the role of weak ties in sharing knowledge across organization subunits', *Administrative Science Quarterly* **44**, 82–111.

Hansen, M.T. (2002), 'Knowledge networks: explaining effective knowledge sharing in multiunit companies', *Organization Science* **13**, 232–48.

Kleinberg, J. (2000), 'Navigation in a Small World', *Nature* **406**, 845.

Krackhardt, D. and J.R. Hanson (1993): 'Informal networks: the company behind the chart', *Harvard Business Review* **71**, 104–11.

Milgram, S. (1967), 'The small-world problem', *Psychology Today* **2**, 60–67.

Moreland, R.L. and L. Argote (2003), 'Transactive memory in dynamic organizations', in R.S. Peterson and E.A. Mannix (eds.), *Leading and Managing People in the Dynamic Organization*, New Jersey: Lawrence Erlbaum Associates, Publishers.

Newman, M.E.J., C. Moore and D.J. Watts (2000), 'Mean-field solution of the small-world network model', *Physical Review Letters* **84**, 3201–04.

Radner, R. (1993), 'The organization of decentralized information processing', *Econometrica* **61**, 1109–1146.

Schein, E.H. (2002), 'Models and tools for stability and change in human systems', *Reflections* **5**, 34–45.

Schein, E.H. (2004), *Organizational Culture and Leadership*, Third Edition, New York: Wiley Publishers.

Schulz M. (2003), 'Pathways of relevance: exploring inflows of knowledge into subunits of multinational corporations', *Organization Science* **14**, 440–59.

van Zandt, T. (1999), 'Real-time decentralized information processing as a model of organizations with boundedly rational agents', *Review of Economic Studies* **66**, 633–658.

Watts, D.J. and S.H. Strogatz (1998), 'Collective dynamics of 'small-world' networks', *Nature* **393**, 440–42.

Wegner, D.M. (1986), 'Transactive memory: A contemporary analysis of the group mind', in B. Mullen and G.R. Goethals (eds.), *Theories of Group Behaviorocial Cognition*, New York: Springer Verlag, pp. 253–76.

Wegner, D.M. (1995): 'A computer network model of human transactive memory', *Social Cognition* **14**, 319–39.

Farsightedly Stable Networks

P. Jean-Jacques Herings, Ana Mauleon and Vincent Vannetelbosch

Paper presented at CTN Workshop 2009. Reprinted from *Games and Economic Behavior*, **67**(2), November 2009, 526–541, with permission from Elsevier.

A set of networks G is pairwise farsightedly stable (i) if all possible farsighted pairwise deviations from any network g ∈ G to a network outside G are deterred by the threat of ending worse off or equally well off, (ii) if there exists a farsighted improving path from any network outside the set leading to some network in the set, and (iii) if there is no proper subset of G satisfying conditions (i) and (ii). A non-empty pairwise farsightedly stable set always exists. We provide a full characterization of unique pairwise farsightedly stable sets of networks. Contrary to other pairwise concepts, pairwise farsighted stability yields a Pareto dominant network, if it exists, as the unique outcome. Finally, we study the relationship between pairwise farsighted stability and other concepts such as the largest pairwise consistent set and the von Neumann-Morgenstern pairwise farsightedly stable set.

1. Introduction

The organization of individual agents into networks and groups or coalitions plays an important role in the determination of the outcome of many social and economic interactions. For instance, networks of personal contacts are important in obtaining information on goods and services, like product information or information about job opportunities. Many commodities are traded through

We would like to thank two anonymous referees for helpful comments. P. Jean-Jacques Herings would like to thank the Netherlands Organization for Scientific Research (NWO) for financial support. Vincent Vannetelbosch and Ana Mauleon are Research Associates of the National Fund for Scientific Research (FNRS). Vincent Vannetelbosch is Associate Fellow of CEREC, Facultés Universitaires Saint-Louis. Financial support from Spanish Ministerio de Educacion y Ciencia under the project SEJ 2006-06309/ECON, support from the Belgian French Community's program Action de Recherches Concertée 03/08-302 and 05/10-331 (UCL) and support of a SSTC grant from the Belgian Federal government under the IAP contract P6/09 are gratefully acknowledged.

networks of buyers and sellers. The partitioning of societies into groups is also important in many contexts, such as the provision of public goods and the formation of alliances, cartels, and federations.

A simple way to analyze the networks that one might expect to emerge in the long run is to examine the requirement that individuals do not benefit from altering the structure of the network. An example of such a condition is the pairwise stability notion defined by Jackson and Wolinsky (1996).[1] Their approach is static and myopic. Individuals are not forward-looking in the sense that they do not forecast how others might react to their actions. For instance, individuals might not add a link that appears valuable to them given the current network, as that might in turn lead to the formation of other links and ultimately lower the payoffs of the original individuals.

A dynamic (but still myopic) network formation process has been recently studied by Jackson and Watts (2002), who have proposed a dynamic process in which individuals form and sever links based on the improvement that the resulting network offers them relative to the current network. This deterministic dynamic process may end at a pairwise stable network or may cycle.

In this paper we address the question which networks one might expect to emerge in the long run when players are farsighted. Since most of the literature considers the case where at most one link is changed at a time, we will also restrict our analysis to network formation processes with this characteristic. This enables us to make the best comparison of our results to those found in the literature. It is straightforward to adapt our concept to more general network formation processes.

We first extend the Jackson and Wolinsky pairwise stability notion to a new set-valued solution concept, called the pairwise myopically stable set. A set of networks G is pairwise myopically stable (i) if all possible myopic pairwise deviations from any network $g \in G$ to a network outside the set are deterred by the threat of ending worse off or equally well off, (ii) if there exists a *myopic improving path* from any network outside the set leading to some network in the set, and (iii) if there is no proper subset of G satisfying conditions (i) and (ii). We show that there is a unique pairwise myopically stable set and that it is equal to the collection of closed cycles. It follows that the pairwise myopically stable set is non-empty and contains all pairwise stable networks.

We then introduce the pairwise farsightedly stable set, to predict which networks may be formed among farsighted players. The definition corresponds to the one of a pairwise myopically stable set with myopic deviations and myopic

[1] There are alternative ways to model network stability. One is to explicitly model a game by which links form and then to solve that game using the concept of Nash equilibrium or one of its refinements. See Aumann and Myerson (1988) and Dutta and Mutuswami (1997). Jackson (2003, 2005) provides surveys of models of network formation.

improving paths replaced by farsighted deviations and farsighted improving paths. A farsighted improving path is a sequence of networks that can emerge when players form or sever links based on the improvement the end network offers relative to the current network. Each network in the sequence differs by one link from the previous one. If a link is added, then the two players involved must both prefer the end network to the current network, with at least one of the two strictly preferring the end network. If a link is deleted, then it must be that at least one of the two players involved in the link strictly prefers the end network.

In contrast to other concepts incorporating farsightedness, we do not only request that all possible pairwise deviations out of the set are deterred by the threat of ending worse off, but also that there exists a farsighted improving path from any network outside the set leading to some network in the set. This property is equivalent to the requirement that networks within the set are robust to perturbations. Perturbations may be due to exogenous forces acting on the network, or simply miscalculations or errors on the part of an individual making an assessment or taking an action.[2]

We show that a pairwise farsightedly stable set always exists and we provide a full characterization of unique pairwise farsightedly stable sets of networks. As a corollary, we give the necessary and sufficient condition such that a unique pairwise farsightedly stable set consisting of a single network exists. We apply these results to examples, such as the criminal network model of Calvó-Armengol and Zenou (2004). We find that in criminal networks with n players, the set consisting of the complete network (where all criminals are linked to each other) is a pairwise farsightedly stable set.

We consider the relationship between farsighted stability and efficiency of networks. We provide conditions under which pairwise farsighted stability singles out a strongly efficient network. We show that if there is a network that Pareto dominates all other networks, then that network is the unique prediction of pairwise farsighted stability. This property does not hold for other pairwise solution concepts.

Finally, we study the relationship between pairwise farsighted stability and other farsighted concepts such as the largest pairwise consistent set, a notion due to Chwe (1994), and the von Neumann-Morgenstern pairwise farsightedly stable set. Under some conditions, a pairwise farsightedly stable set is a subset of the set of pairwise stable networks, which in turn is a subset of the largest pairwise consistent set. We show that any von Neumann-Morgenstern pairwise farsightedly

[2] Jackson and Watts (2002) use improving paths as the foundation for a stochastic analysis, where in addition to intended changes in the network, unintended mutations or errors are introduced. However, in their definition of improving path it is assumed that players behave myopically: all a player needs to know is whether adding or deleting a given link is directly beneficial to him or her under the current circumstances.

stable set is also a pairwise farsightedly stable set. Under some conditions, also the reverse statement holds. By means of examples we show that pairwise farsightedly stable sets have no relationship to either largest pairwise consistent sets or pairwise myopically stable sets.

Although the literature on stability in networks is well established and growing (see Jackson, 2005), the literature on farsighted stability is still in its infancy. Page et al. (2005) address the issue of farsighted stability in network formation by extending Chwe's (1994) result on the nonemptiness of farsightedly consistent sets. In order to demonstrate the existence of farsightedly consistent directed networks, they provide a new framework that extends the standard notion of a directed network and also introduces the notion of a supernetwork. A supernetwork specifies how the different directed networks are connected via coalitional moves and coalitional preferences, and thus provides a network representation of agent preferences and the rules governing network formation. A supernetwork is equivalent to the social environment studied by Chwe (1994), when the set of outcomes is replaced by the set of directed networks. Given the rules governing network formation and agents' preferences as represented via the supernetwork, a directed network (i.e., a particular node in the supernetwork) is said to be farsightedly consistent if no agent or coalition of agents is willing to alter the network (via the addition, subtraction, or replacement of arcs) in fear that such an alteration might induce further network alterations by other agents or coalitions that in the end leave the initially deviating agent or coalition no better off, and possibly worse off. They have shown that for any supernetwork corresponding to a given collection of directed networks, the set of farsightedly consistent networks is non-empty; see also Page and Wooders (2005).

Dutta et al. (2005) have studied a model of dynamic network formation where individuals are farsighted and evaluate the desirability of a move in terms of its consequences on the entire discounted stream of payoffs. Contrary to ours, their model is in spirit closer to non-cooperative game theory. They show that a Markovian equilibrium process of network formation exists and they provide two conditions, link monotonicity and increasing returns to link creation, each of which guarantees that there is some equilibrium at which the complete graph is reached in the limit from all initial networks. They also show that there are valuation structures in which the process will not converge to any efficient network for any equilibrium strategy profile. This can be viewed as the dynamic counterpart of the conflict between static stability and efficiency demonstrated by Jackson and Wolinsky (1996), a conflict that is also confirmed by our results. Dutta et al. (2005) provide an example where there is a network that Pareto dominates all other networks, but which is not reached in equilibrium. In our framework, such an example is not possible. If there is a network that Pareto dominates all other networks, then that network is the unique prediction of pairwise farsighted stability. Other approaches to farsightedness in network formation are suggested by the work of Xue (1998), Herings et al. (2004), and Mauleon and Vannetelbosch (2004).

The paper is organized as follows. In Section 2 we introduce some notations and basic properties and definitions for networks. In Section 3 we define the notion of pairwise myopically stable set of networks. In Section 4 we define the notion of pairwise farsightedly stable set of networks and we characterize it in Section 5. In Section 6 we consider the symmetric connections model and the co-author model. We look at the relationship between farsighted stability and efficiency of networks in Section 7. In Section 8 and Section 9 we analyze, respectively, the relationship with the von Neumann-Morgenstern pairwise farsightedly stable set and the largest pairwise consistent set. Finally, in Section 10 we conclude.

2. Networks

Let $N = \{1, \dots, n\}$ be the finite set of players who are connected in some network relationship. The network relationships are reciprocal and the network is thus modeled as a non-directed graph. Individuals are the nodes in the graph and links indicate bilateral relationships between individuals. Thus, a network g is simply a list of which pairs of individuals are linked to each other. We write $ij \in g$ to indicate that i and j are linked under the network g. Let g^N be the collection of all subsets of N with cardinality 2, so g^N is the complete network. The set of all possible networks or graphs on N is denoted by $\mathbb{G}$ and consists of all subsets of g^N. The network obtained by adding link ij to an existing network g is denoted $g + ij$ and the network that results from deleting link ij from an existing network g is denoted $g - ij$. For any network g, let $N(g) = \{i \mid \text{there is } j \text{ such that } ij \in g\}$ be the set of players who have at least one link in the network g. A path in a network $g \in \mathbb{G}$ between i and j is a sequence of players $i_1, \dots, i_K$ such that $i_k i_{k+1} \in g$ for each $k \in \{1, \dots, K - 1\}$ with $i_1 = 1$ and $i_K = j$. A non-empty network $h \subseteq g$ is a component of g, if for all $i \in N(h)$ and $j \in N(h)\setminus\{i\}$, there exists a path in h connecting i and j, and for any $i \in N(h)$ and $j \in \mathrm{N}(g)$, $ij \in g$ implies $ij \in h$. The set of components of g is denoted by $C(g)$. Knowing the components of a network, we can partition the players into groups within which players are connected. Let $\prod(g)$ denote the partition of N induced by the network g.[3]

A value function is a function $v: \mathbb{G} \to \mathbb{R}$ that keeps track of how the total societal value varies across different networks. The set of all possible value functions is denoted by $\mathcal{V}$. An allocation rule is a function $Y: \mathbb{G} \times \mathcal{V} \to \mathbb{R}^N$ that keeps track of how the value is allocated or distributed among the players forming a network. It satisfies $\sum_{i \in N} Y_i(g, v) = v(g)$ for all v and g.

[3] Throughout the paper we use the notation $\subseteq$ for weak inclusion and $\subsetneq$ for strict inclusion. Finally, # will refer to the notion of cardinality.

Jackson and Wolinsky (1996) have proposed a number of basic properties of value and allocation functions. A value function is *component additive* if $v(g) = \sum_{h \in C(g)} v(h)$ for all $g \in \mathbb{G}$. Component additive value functions are the ones for which the value of a network is the sum of the value of its components. An allocation rule Y is *component balanced* if for any component additive $v \in \mathcal{V}$, $g \in \mathbb{G}$, and $h \in C(g)$, we have $\sum_{i \in N(h)} Y_i(h, v) = v(h)$. Component balancedness only puts conditions on Y for v's that are component additive, so Y can be arbitrary otherwise. Given a permutation of players π and any $g \in \mathbb{G}$, let $g^{\pi} = \{\pi(i)\pi(j) \mid ij \in g\}$. Thus, g^{π} is a network that is identical to g up to a permutation of the players. A value function is *anonymous* if for any permutation π and any $g \in \mathbb{G}$, $v(g^{\pi}) = v(g)$. Given a permutation π, let v^{π} be defined by $v^{\pi}(g) = v(g^{\pi^{-1}})$ for each $g \in \mathbb{G}$. An allocation rule Y is *anonymous* if for any $v \in \mathcal{V}$, $g \in \mathbb{G}$, and permutation π, we have $Y_{\pi(i)}(g^{\pi}, v^{\pi}) = Y_i(g, v)$.[4]

An allocation rule that is component balanced and anonymous is the *componentwise egalitarian allocation rul*e. For a component additive v and network g, the componentwise egalitarian allocation rule Y^{ce} is such that for any $h \in C(g)$ and each $i \in N(h)$, $Y_i^{ce}(g, v) = v(h)/\#N(h)$. For a v that is not component additive, $Y^{ce}(g, v) = v(g)/n$ for all g; thus, Y^{ce} splits the value $v(g)$ equally among all players if v is not component additive.

In evaluating societal welfare, we may take various perspectives. A network g is *Pareto efficient* relative to v and Y if there does not exist any $g' \in \mathbb{G}$ such that $Y_i(g',v) \geq Y_i(g, v)$ for all i with at least one strict inequality. A network $g \in \mathbb{G}$ is *strongly efficient* relative to v if $v(g) \geq v(g')$ for all $g' \in \mathbb{G}$. This is a strong notion of efficiency as it takes the perspective that value is fully transferable.

The network-theoretic literature uses two different notions of deviation by a coalition. *Pairwise deviations* (Jackson and Wolinsky, 1996) are deviations involving a single link at a time. Moreover, link addition is bilateral (two players that would be involved in the link must agree to adding the link), link deletion is unilateral (at least one player involved in the link must agree to deleting the link), and network changes take place one link at a time. *Coalitionwise deviations* (Jackson and van den Nouweland, 2005) are deviations involving several links and some group of players at a time. Link addition is bilateral, link deletion is unilateral, and multiple link changes can take place at a time. Whether a pairwise deviation or a coalitionwise deviation makes more sense will depend on the setting within which network formation takes place.

We will restrict our analysis to pairwise deviations. A simple way to analyze the networks that one might expect to emerge in the long run is to examine the requirement that agents do not benefit from altering the structure of the network.

[4] Anonymous value functions are those such that the architecture of a network matters, but not the labels of individuals. Anonymity of an allocation rule requires that if only the labels of the agents change and the value generated by networks changes in an exactly corresponding fashion, then the allocation only changes according to the relabeling.

A weak version of such a condition is the pairwise stability notion defined by Jackson and Wolinsky (1996). A network is pairwise stable if no player benefits from severing one of their links and no other two players benefit from adding a link between them, with one benefiting strictly and the other at least weakly. Formally, a network g is pairwise stable with respect to value function v and allocation rule Y if

(i) for all $ij \in g$, $Y_i(g, v) \geq Y_i(g - ij, v)$ and $Y_j(g, v) \geq Y_j(g - ij, v)$, and
(ii) for all $ij \notin g$, if $Y_i(g, v) < Y_i(g + ij, v)$ then $Y_j(g, v) > Y_j(g + ij, v)$.

We say that g' is adjacent to g if $g' = g + ij$ or $g' = g - ij$ for some ij. A network g' defeats g if either $g' = g - ij$ and $Y_i(g', v) > Y_i(g, v)$ or $Y_j(g', v) > Y_j(g, v)$, or if $g' = g + ij$ with $Y_i(g', v) \geq Y_i(g, v)$ and $Y_j(g', v) \geq Y_j(g, v)$ with at least one inequality holding strictly. Pairwise stability is equivalent to the statement of not being defeated by another network.[5]

3. Pairwise Myopically Stable Sets of Networks

Pairwise stable networks do not always exist. Following Jackson and Watts (2002), we introduce the notion of myopic improving path. A myopic improving path is a sequence of networks that can emerge when players form or sever links based on the improvement the resulting network offers relative to the current network. Each network in the sequence differs by one link from the previous one. If a link is added, then the two players involved must both prefer the resulting network to the current network, with at least one of the two strictly preferring the resulting network. If a link is deleted, then it must be that at least one of the two players involved in the link strictly prefers the resulting network.

Definition 1. A myopic improving path from a network g to a network $g' \neq g$ is a finite sequence of graphs $g_1, \ldots, g_K$ with $g_1 = g$ and $g_K = g'$ such that for any $k \in \{1, \ldots, K - 1\}$ either:

(i) $g_{k+1} = g_k - ij$ for some ij such that $Y_i(g_{k+1}, v) > Y_i(g_k, v)$ or $Y_j(g_{k+1}, v) > Y_j(g_k, v)$, or
(ii) $g_{k+1} = g_k + ij$ for some ij such that $Y_i(g_{k+1}, v) > Y_i(g_k, v)$ and $Y_j(g_{k+1}, v) \geq Y_j(g_k, v)$.

A myopic improving path is a sequence of adjacent networks that might be observed in a dynamic process where players are adding and deleting links, one at

[5] Jackson and van den Nouweland (2005) have proposed a refinement of pairwise stability where coalitionwise deviations are allowed: the strongly stable networks. A strongly stable network is a network which is stable against changes in links by any coalition of individuals. Strongly stable networks are Pareto efficient and maximize the overall value of the network if the value of each component of a network is allocated equally among the members of that component.

a time. If there exists a myopic improving path from g to g', then we write $g \mapsto g'$. For a given network g, let $M(g) = \{g' \in \mathbb{G} \mid g \mapsto g'\}$. This is the set of networks that can be reached by a myopic improving path from g. Thus, $g \mapsto g'$ means that g' is the endpoint of at least one myopic improving path from g. Notice that if g is pairwise stable then $M(g) = \varnothing$.

It is well-known that there are many allocation rules and value functions for which pairwise stable networks do not exist. We therefore consider a set-valued solution concept that captures the pairwise stability notion, called the pairwise myopically stable set of networks.

Definition 2. A set of networks $G \subseteq \mathbb{G}$ is pairwise myopically stable with respect v and Y if

(i) $\forall g \in G$,

(ia) $\forall ij \notin g$ such that $g + ij \notin G$, $(Y_i(g + ij, v), Y_j(g + ij, v)) = (Y_i(g, v), Y_j(g, v))$ or $Y_i(g + ij, v) < Y_i(g, v)$ or $Y_j(g + ij, v) < Y_j(g, v)$,

(ib) $\forall ij \in g$ such that $g - ij \notin G$, $Y_i(g - ij, v) \leq Y_i(g, v)$ and $Y_j(g - ij, v) \leq Y_j(g, v)$,

(ii) $\forall g' \in \mathbb{G} \setminus G$, $M(g') \cap G \neq \varnothing$,

(iii) $\nexists G' \subsetneq G$ such that G' satisfies conditions (ia), (ib), and (ii).

Conditions (ia) and (ib) in Definition 2 capture deterrence of external deviations. In condition (ia) the addition of a link ij to a network $g \in G$ that leads to a network outside G is deterred because the two players involved do not prefer the resulting network to network g. Condition (ib) is a similar requirement, but then for the case where a link is severed. Condition (ii) requires external stability. External stability asks for the existence of a myopic improving path from any network outside G leading to some network in G. Condition (ii) implies that if a set of networks is pairwise myopically stable, it is non-empty. Notice that the set $\mathbb{G}$ (trivially) satisfies conditions (ia), (ib), and (ii) in Definition 2. This motivates condition (iii), the minimality condition.

Jackson and Watts (2002) define the notion of a closed cycle. A set of networks C is a *cycle* if for any $g \in C$ and $g' \in C \setminus \{g\}$, there exists a myopic improving path connecting g to g'. A cycle C is a maximal cycle if it is not a proper subset of a cycle. A cycle C is a *closed cycle* if no network in C lies on a myopic improving path leading to a network that is not in C. A closed cycle is necessarily a maximal cycle. If g is a pairwise stable network, then trivially $\{g\}$ is a closed cycle.

Lemma 1. *For every $g \in \mathbb{G}$, either g is pairwise stable or there is a closed cycle C such that $C \subseteq M(g)$.*

Proof. Consider a network $g \in \mathbb{G}$ that is not pairwise stable. Construct a sequence of networks $g_1, \ldots, g_k$ such that $g_1 = g$, $g_{j+1} \in M(g_j)$ for $j = 1, \ldots, k - 1$, and either

$M(g_k) = \varnothing$ or $g_k = g_j$ for some $j < k$. In the former case, g_k is pairwise stable and by transitivity of M, $\{g_k\} \subseteq M(g)$. In the latter case, let C^1 be a maximal cycle containing $\{g_j, \ldots, g_{k-1}\}$. Either C^1 is a closed cycle and we are done, or it has a myopic improving path going out of it, leading to a new maximal cycle C^2. Repeating this argument we reach after a finite number of steps a maximal cycle without myopic improving paths going out of it, i.e. a closed cycle. ■

Lemma 1 confirms the result of Jackson and Watts (2002) that for any value function and allocation rule there exists at least one closed cycle of networks. The next result claims that there is a unique pairwise myopically stable set of networks. It contains all networks that belong to a closed cycle.

Theorem 1. *The set of networks consisting of all networks that belong to a closed cycle is the unique pairwise myopically stable set.*

Proof. Let G be the set consisting of all networks that belong to a closed cycle. We show that G satisfies conditions (i), (ii), and (iii). Obviously, G satisfies condition (i). The set G also satisfies condition (ii). Indeed, consider some $g' \notin G$. By Lemma 1, either g' is pairwise stable or there is a closed cycle C such that $C \subseteq M(g')$. The former case contradicts $g' \notin G$. The latter case implies $g \in M(g')$ for some $g \in G$, i.e. implies condition (ii). Suppose G does not satisfy condition (iii). Let $G' \subsetneq G$ satisfy conditions (i) and (ii). Let g be a network that belongs to G but not to G'. If g is pairwise stable, then there is no $g' \in G'$ such that $g' \in M(g)$, so G' violates condition (ii). The network g is therefore part of a closed cycle C consisting of at least two networks. Moreover, for every $\bar{g} \in C$ it holds that $g \in M(\bar{g})$. If $C \cap G' = \varnothing$, we violate condition (ii). If $C \cap G'$ contains a network $\bar{g}$, then $g \in M(\bar{g})$ implies that condition (i) is violated. It follows that a set $G' \subsetneq G$ satisfying conditions (i) and (ii) does not exist, so G is a pairwise myopically stable set.

Next we show that if a set of networks G satisfies conditions (i)-(iii), then G consists of all networks that belong to a closed cycle. If G does not contain a pairwise stable network, then we have a contradiction to condition (ii). If G does not contain some network that belongs to a closed cycle with at least two elements, then we have a contradiction to condition (i). It follows that G contains all networks that belong to a closed cycle. condition (iii) together with the first part of the proof now yields that G cannot contain any other networks. ■

Sengupta and Sengupta (1994) define an indirect dominance relation for transferable utility games in coalition structure that is analogous to our notion of an improving path. Following them, we can define a network g to be *viable* if for every network $g' \in M(g)$ it holds that $g \in M(g')$. It is easily verified that any viable network g belongs to a closed cycle. Indeed, if g is viable, then $M(g)$ is a closed cycle, and vice versa. We can therefore rephrase Theorem 1 as the statement that the unique pairwise myopically stable set coincides with the set of viable networks.

4. Pairwise Farsightedly Stable Sets of Networks

We start this section with an example that shows the limitations of the pairwise myopically stable set and that motivates the incorporation of an appropriate notion of farsightedness.

EXAMPLE 1 – *Criminal networks*.[6] Each player is a criminal. If two players are connected, then they are part of the same criminal network. Each group of connected criminals has a positive probability of winning the loot. The loot is divided among the connected criminals based on the network architecture. Criminal i's payoff is given by

$$\begin{aligned} Y_i(g) &= p_i(g) \cdot [y_i(g) \cdot (1 - \phi)] + (1 - p_i(g)) \cdot y_i(g) \\ &= y_i(g) \cdot [1 - p_i(g) \cdot \phi], \end{aligned}$$

where $y_i(g)$ is i's expected share of the loot, $p_i(g)$ is i's probability of being caught, and $\phi > 0$ is the penalty rate.[7] Beside being competitors in the crime market, criminals may also benefit from having criminal mates. It is assumed that (i) the bigger the group of connected criminals, the higher its probability of getting the loot, and (ii) the higher the number of links a criminal has, the lower his individual probability of being caught. Suppose the probability of being caught for criminal i is simply given by

$$p_i(g) = \frac{n-1-n_i}{n},$$

with n_i the number of links criminal i has.[8] For any group $S \in \prod(g)$ of connected criminals, let $\overline{n}(S) = \max_{i \in S}[n_i]$. A criminal i that is part of a group $S \in \prod(g)$ expects a share of the loot $B > 0$ given by

$$y_i(g) = \frac{|S|}{n} \cdot \alpha_i(g) \cdot B,$$

[6] This is a simplified version of Calvó-Armengol and Zenou (2004) where, in addition to forming links with criminal mates, criminals choose their level of criminal activities and whether or not to be involved in criminal activities.

[7] The value function v is simply $v(g) = \sum_{i \in N} Y_i(g)$. Since v is fixed, we omit it in the notation of $Y_i(v, g)$.

[8] This assumption captures the idea that delinquents learn from other criminals belonging to the same network how to commit crime in a more efficient way by sharing the know-how about the technology of crime (see Calvó-Armengol and Zenou, 2004)

where $|S|/n$ is the probability that group S will win the loot, and $\alpha_i(g)$ is the share of the loot criminal $i \in S$ would obtain, which is given by

$$\alpha_i(g) = \begin{cases} \frac{1}{\#\{j \in S \mid n_j = \overline{n}(S)\}} & \text{if } n_i = \overline{n}(S), \\ 0 & \text{otherwise.} \end{cases}$$

That is, within each component, the criminal who has the highest number of links gets the loot. If two or more criminals have the highest number of links, then they share the loot equally among them. In Figure 1 we have depicted the 3-player case with all payoffs in 1/9-th's. For $\phi < 3/2$, both the partial networks (g_1, g_2, g_3) and the complete network (g_7) are pairwise stable networks. There are four closed cycles, $\{g_1\}$, $\{g_2\}$, $\{g_3\}$, and $\{g_7\}$. The pairwise myopically stable set of networks is therefore given by $\{g_1, g_2, g_3, g_7\}$. For $\phi \geq 3/2$, the complete network g_7 is the only pairwise stable network. There is only one closed cycle, $\{g_7\}$, which is therefore also the pairwise myopically stable set of networks.

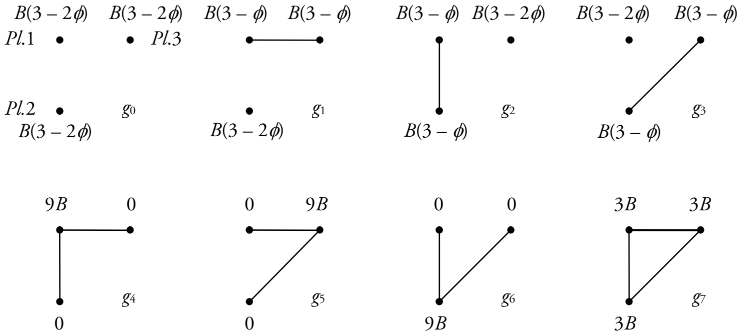


Figure 1. Criminal networks

Take some strictly positive ϕ smaller than 3/2 in Example 1. The partial networks g_1, g_2, and g_3 are pairwise stable. Notice that two links have to be added to a partial network g_1, g_2, or g_3 to form the complete network g_7. Farsighted players may decide to add one link to a network like g_1, g_2, or g_3, accepting a loss, in the expectation that a further link will be added to form the complete network. A *farsighted improving path* is a sequence of networks that can emerge when players form or sever links based on the improvement the end network offers relative to the current network. Each network in the sequence differs by one link from the previous one. If a link is added, then the two players involved must both prefer the end network to the current network, with at least one of the two strictly preferring the end network. If a link is deleted, then it must be that at least one of the two players involved in the link strictly prefers the end network. We now introduce the formal definition of a farsighted improving path.

Definition 3. A farsighted improving path from a network g to a network $g' \neq g$ is a finite sequence of graphs $g_1, \dots, g_K$ with $g_1 = g$ and $g_K = g'$ such that for any $k \in \{1, \dots, K-1\}$ either:

(i) $g_{k+1} = g_k - ij$ for some ij such that $Y_i(g_K, v) > Y_i(g_k, v)$ or $Y_j(g_K, v) > Y_j(g_k, v)$, or
(ii) $g_{k+1} = g_k + ij$ for some ij such that $Y_i(g_K, v) > Y_i(g_k, v)$ and $Y_j(g_K, v) \geq Y_j(g_k, v)$.

If there exists a farsighted improving path from g to g', then we write $g \mapsto g'$. For a given network g, let $F(g) = \{g' \in \mathbb{G} \mid g \mapsto g'\}$. This is the set of networks that can be reached by a farsighted improving path from g. Thus, $g \mapsto g'$ means that g' is the endpoint of at least one farsighted improving path from g. Notice that $F(g)$ may contain many networks and that a network $g' \in F(g)$ might be the endpoint of several farsighted improving paths starting in g. Since we are interested in stability of networks, there will be no need to specify on which particular path players eventually agree. Rather $F(g)$ represents the networks that could possibly be reached by farsighted players when starting in g, and our concept of stability takes these possible end networks into account in a way that we will make precise in Definition 4.

The notion of farsightedness is relevant whenever payoffs can only be reaped after some stable network has formed, or when players are sufficiently patient so that they can safely ignore the payoffs that they obtain before a stable network settles down. It lies at the other end of the spectrum than myopia, where only immediate payoffs count. An intermediate (but difficult) approach is the one of Dutta et al. (2005), where the entire discounted stream of payoffs matters.

Suppose in Example 1 with ϕ smaller than 3/2 that the starting network g is a partial network like g_1, g_2, or g_3. Then, from g no myopic improving path results in the complete network. The problem is that the isolated player will loose from making a link with any of the other players. However, there are farsighted improving paths that go to the complete network. An example of the sequence of graphs on a farsighted improving path is (g_1, g_4, g_7) when starting in g_1. Similar farsighted improving paths exist starting in any of the other partial networks. Examples of other farsighted improving paths starting in g_1 and ending in g_7 are (g_1, g_4, g_7), (g_1, g_5, g_7), or even $(g_1, g_0, g_2, g_6, g_7)$. Moreover, from any $g \neq g_7$ there is a farsighted improving path going to g_7. Thus, we observe that the partial networks are pairwise stable, but not stable when players are farsighted. The complete network on the other hand is not only pairwise stable. It is also stable when players are farsighted.

We now introduce a new solution concept, the pairwise farsightedly stable set. The definition corresponds to the one of a pairwise myopically stable set with myopic deviations replaced by farsighted deviations. It is obtained by requiring the deterrence of farsighted external deviations, farsighted external stability, and minimality. More precisely, a set of networks G is pairwise farsightedly stable if (i) all possible *pairwise deviations* from any network $g \in G$ to a network outside G are deterred by a credible threat of ending worse off or equally well off, (ii) there exists a farsighted improving path from any network outside the set leading to some network in the set, and (iii) there is no proper subset of G satisfying conditions (i) and (ii). Formally, pairwise farsightedly stable sets are defined as follows.

Definition 4. A set of networks $G \subseteq \mathbb{G}$ is pairwise farsightedly stable with respect v and Y if

(i) $\forall g \in G$,

(ia) $\forall ij \notin g$ such that $g + ij \notin G$, $\exists g' \in F(g + ij) \cap G$ such that $(Y_i(g', v), Y_j(g', v)) = (Y_i(g, v), Y_j(g, v))$ or $Y_i(g', v) < Y_i(g, v)$ or $Y_j(g', v) < Y_j(g, v)$,

(ib) $\forall ij \in g$ such that $g - ij \notin G$, $\exists g', g'' \in F(g - ij) \cap G$ such that $Y_i(g', v) \leq Y_i(g, v)$ and $Y_j(g'', v) \leq Y_j(g, v)$,

(ii) $\forall g' \in \mathbb{G} \backslash G$, $F(g') \cap G \neq \emptyset$.

(iii) $\nexists G' \subsetneq G$ such that G' satisfies conditions (ia), (ib), and (ii).

Condition (ia) in Definition 4 captures that adding a link ij to a network $g \in G$ that leads to a network outside of G, is deterred by the threat of ending in g'. Here g' is such that there is a farsighted improving path from $g + ij$ to g'. Moreover, g' belongs to G, which makes g' a credible threat. Condition (ib) is a similar requirement, but then for the case where a link is severed. Condition (ii) in Definition 4 requires external stability and implies that the networks within the set are robust to perturbations. From any network outside G there is a farsightedly stable path leading to some network in G.[9] Condition (ii) implies that if a set of networks is pairwise farsightedly stable, it is non-empty. Notice that the set $\mathbb{G}$ (trivially) satisfies conditions (ia), (ib), and (ii) in Definition 4. This motivates the requirement of a minimality condition, namely condition (iii).

Theorem 2. *A pairwise farsightedly stable set of networks exists.*

Proof. Notice that $\mathbb{G}$ satisfies conditions (i) and (ii). Let us proceed by contradiction. Assume that there does not exist any set of networks $G \subseteq \mathbb{G}$ that is pairwise farsightedly stable. This means that for any $G^0 \subseteq \mathbb{G}$ that satisfies conditions (i) and (ii) in Definition 4, we can find a proper subset G^1 that satisfies conditions (i) and (ii). Iterating this reasoning we can build an infinite decreasing sequence $\{G^k\}_{k \geq 0}$ of subsets of $\mathbb{G}$ satisfying conditions (i) and (ii). But since $\mathbb{G}$ has finite cardinality, this is not possible. ■

[9] There are some random dynamic models of network formation that are based on incentives to form links such as Watts (2001), Jackson and Watts (2002), and Tercieux and Vannetelbosch (2006). These models aim to use the random process to select from the set of pairwise stable networks.

We show next that in Example 1 with $n = 3$, the set consisting of the complete network is the unique pairwise farsightedly stable set whatever the fine ϕ.

We consider first the case $\phi < 3/2$. It can be verified that $F(g_0) = \{g_1, g_2, g_3, g_7\}$, $F(g_1) = \{g_2, g_3, g_7\}$, $F(g_2) = \{g_1, g_3, g_7\}$, $F(g_3) = \{g_1, g_2, g_7\}$, $F(g_4) = \{g_1, g_2, g_3, g_7\}$, $F(g_5) = \{g_1, g_2, g_3, g_7\}$, $F(g_6) = \{g_1, g_2, g_3, g_7\}$, and $F(g_7) = \emptyset$. Notice that the analysis of farsighted improving paths can be intricate. The only way to go from g_1 to g_2 is via g_4. At the same time it holds that $g_4 \notin F(g_1)$. Indeed, players 1 and 2 make a link to go from g_1 to the intermediate network g_4 in the anticipation that player 3 will subsequently delete his link with player 1. To go from g_1 to the terminal network g_4 is a strict deterioration for players 2 and 3. The only thing player 1 can do is to sever his link with player 2, which leads to g_0. This is not helpful for player 1, since once at g_0 he is still the only one that is better off at g_4 compared to g_0, and there is nothing that he can do anymore.

We show next that $\{g_7\}$ is pairwise farsightedly stable. Since $g_7 \in \bigcap_{g \in \mathbb{G}\setminus\{g_7\}} F(g)$, condition (ii) of the definition is clearly satisfied. Moreover, condition (i) is satisfied, since any deviation from g_7 may lead back to g_7. Clearly, $\{g_7\}$ is minimal, so condition (iii) is satisfied too.

There are no other pairwise farsightedly stable sets. Since $F(g_7) = \emptyset$, condition (ii) implies that g_7 belongs to any pairwise farsightedly stable set. Since $\{g_7\}$ is pairwise farsightedly stable, using condition (iii) it follows that $\{g_7\}$ is the only pairwise farsightedly stable set.

Take now $\phi \geq 3/2$. For $3/2 \leq \phi \leq 3$ we have $F(g_0) = \{g_1, g_2, g_3, g_4, g_5, g_6, g_7\}$, $F(g_1) = \{g_4, g_5, g_7\}$, $F(g_2) = \{g_4, g_6, g_7\}$, $F(g_3) = \{g_5, g_6, g_7\}$, $F(g_4) = \{g_7\}$, $F(g_5) = \{g_7\}$, $F(g_6) = \{g_7\}$, and $F(g_7) = \emptyset$. For $\phi > 3$ we have $F(g_0) = \{g_1, g_2, g_3, g_4, g_5, g_6, g_7\}$, $F(g_1) = \{g_4, g_5, g_6, g_7\}$, $F(g_2) = \{g_4, g_5, g_6, g_7\}$, $F(g_3) = \{g_4, g_5, g_6, g_7\}$, $F(g_4) = \{g_7\}$, $F(g_5) = \{g_7\}$, $F(g_6) = \{g_7\}$, and $F(g_7) = \emptyset$. So, $g_7 \in \bigcap_{g \in \mathbb{G}\setminus\{g_7\}} F(g)$ for $\phi \geq 3/2$. Since $F(g_7) = \emptyset$, we can use the same arguments as in the case $\phi < 3/2$ and can therefore conclude that $\{g_7\}$ is the unique pairwise farsightedly stable set.

5. Characterizations of Pairwise Farsightedly Stable Sets

The next theorem provides an easy to verify condition for a set G to be pairwise farsightedly stable.

Theorem 3. *If for every $g' \in \mathbb{G}\setminus G$ we have $F(g') \cap G \neq \emptyset$ and for every $g \in G$, $F(g) \cap G = \emptyset$, then G is a pairwise farsightedly stable set.*

Proof. Condition (ii) is trivially satisfied.

Suppose condition (i) is not satisfied. Then there is $g \in G$ and a deviation to $g' \notin G$ such that every $g'' \in F(g') \cap G$ defeats g. In particular, it then follows that $g'' \in F(g)$, a contradiction, since by assumption there is no $g'' \in G$ with that property. Consequently, condition (i) holds.

To verify condition (iii), suppose there is a proper subset G' of G that satisfies conditions (i) and (ii). Let g be in G but not in G'. Then $F(g) \cap G' \subseteq F(g) \cap G = \varnothing$, where the equality follows by the assumption in the theorem. It follows that G' violates condition (ii), leading to a contradiction. We have shown that G is minimal. ■

Later on in the paper, we will show by means of Example 3 that Theorem 3 cannot be extended to an 'if and only if' statement. The 'if and only if' statement is true, however, when restricting the scope of the theorem to sets consisting of a single network.

Theorem 4. *The set $\{g\}$ is a pairwise farsightedly stable set if and only if for every $g' \in \mathbb{G}\backslash\{g\}$ we have $g \in F(g')$.*

Proof. If $\{g\}$ is a pairwise farsightedly stable set, then by condition (ii) in Definition 4 it follows that $g \in F(g')$ for every $g' \in \mathbb{G}\backslash\{g\}$.

Now suppose that for every $g' \in \mathbb{G}\backslash\{g\}$ we have $g \in F(g')$. Condition (ii) is trivially satisfied. Since $g \in F(g + ij)$ and $g \in F(g - ij)$, conditions (ia) and (ib) hold. Finally, condition (iii) is satisfied because $\{g\}$ is a singleton. ■

Theorem 4 tells us that $\{g\}$ is a pairwise farsightedly stable set if and only if there exists a farsighted improving path from any network leading to g. Condition (iii) of the definition implies that if $\{g\}$ is a pairwise farsightedly stable set, then g does not belong to any other pairwise farsightedly stable set. But there may be pairwise farsightedly stable sets not containing g.

The next result provides a full characterization for unique pairwise farsightedly stable sets.

Theorem 5. *The set G is the unique pairwise farsightedly stable set if and only if $G = \{g \in \mathbb{G} \mid F(g) = \varnothing\}$ and for every $g' \in \mathbb{G}\backslash G$, $F(g') \cap G \neq \varnothing$.*

Proof. ($\Leftarrow$) Condition (ii) of Definition 4 is trivially satisfied. Suppose condition (i) is not satisfied. Then there is $g \in G$ and $ij \notin g$ such that $g + ij \notin G$ and for every $g' \in F(g + ij) \cap G$ it holds that $(Y_i(g', v), Y_j(g', v)) > (Y_i(g, v),Y_j(g, v))$, or there is $g \in G$ and $ij \in g$ such that $g - ij \notin G$ and for every $g' \in F(g + ij) \cap G$ it holds that $Y_i(g', v) > Y_i(g, v)$. In both cases it follows that $g' \in F(g)$, a contradiction, since by assumption $F(g) = \varnothing$. Consequently, condition (i) holds. Since for every $g \in G$, $F(g) = \varnothing$, by condition (ii) it holds that G is a subset of any pairwise farsightedly stable set. It then follows from condition (iii) that G is the unique pairwise farsightedly stable set.

($\Rightarrow$) Condition (ii) yields that for every $g' \in \mathbb{G}\backslash G$, $F(g') \cap G \neq \varnothing$. From this it also follows that every g with $F(g) = \varnothing$ belongs to G. It remains to be shown that for every $g \in G$, $F(g) = \varnothing$. Suppose not. Let g^* and g' be such that $g^* \in G$ and $g' \in F(g^*)$. Consider $G' = \{g'\} \cup \{g \in \mathbb{G} \mid g' \notin F(g)\}$. Notice that $g^* \notin G'$ and that for any $g \notin G'$ we have that $g' \in F(g)$.

Claim. G' satisfies conditions (i) and (ii).

Since for any $g \notin G'$ we have that $g' \in F(g)$, condition (ii) is satisfied. Consider any pairwise deviation from g' to $g'' \notin G'$. By construction of G', $g' \in F(g'')$ and the deviation is deterred. Consider any pairwise deviation from any $g^0 \in G'\backslash\{g'\}$ to some $g'' \notin G'$. Suppose that all $g \in F(g'') \cap G'$ are preferred by the player(s) initially deviating from g^0. Then it follows that $F(g'') \cap G' \subseteq F(g^0)$. By definition of G', $g' \in F(g'')$, so $g' \in F(g'') \cap G' \subseteq F(g^0)$, contradicting $g' \notin F(g^0)$ for any $g^0 \in G'\backslash\{g'\}$. Consequently, all pairwise deviations from $g^0 \in G'\backslash\{g'\}$ are deterred. Since we already showed that pairwise deviations from g' are deterred too, the set G' satisfies condition (i).

Finally, if G' satisfies condition (iii), then G' is a pairwise farsightedly stable set, a contradiction to the uniqueness of G. If G' does not satisfy condition (iii), then, following the reasoning in the proof of Theorem 2, there is a proper subset G'' of G' satisfying conditions (i), (ii), and (iii). Since $g^* \in G\backslash G''$, it holds that $G \neq G''$ and we obtain a contradiction to the uniqueness of G. ■

Theorem 5 implies that if G is the unique pairwise farsightedly stable set and the network g belongs to G, then $F(g) = \varnothing$, which implies that g is pairwise stable. Thus, pairwise farsighted stability is a refinement of pairwise stability when there is a unique pairwise farsightedly stable set.

From Theorem 5 we obtain the following corollary that provides the necessary and sufficient conditions such that there is a unique pairwise farsightedly stable set consisting of a single network.

Corollary 1. *The set $\{g\}$ is the unique pairwise farsightedly stable set if and only if for every $g' \in \mathbb{G}\backslash\{g\}$ we have $g \in F(g')$ and $F(g) = \varnothing$.*

If for every $g' \in \mathbb{G}\backslash\{g\}$ we have $g \in F(g')$, then by Theorem 4 $\{g\}$ is a pairwise farsightedly stable set. If, moreover, $F(g) = \varnothing$, then $\{g\}$ is the unique pairwise farsightedly stable set by Corollary 1. If, on the other hand, $F(g) \neq \varnothing$, then there is another pairwise farsightedly stable set by Corollary 1.

Using Theorem 4 we prove that in the example of criminal networks with n players, the complete network $\{g^N\}$ is a pairwise farsightedly stable set.

Proposition 1. In the criminal networks model, the set $\{g^N\}$ is a pairwise farsightedly stable set.

Proof. We show that for every $g \in \mathbb{G}\backslash\{g^N\}$ we have $g^N \in F(g)$. (1) We show that from any $g \neq g^N$ there is always a player who wants to delete a link looking forward to g^N. This enables us to build a sequence of networks where at each step some player (who is looking forward to g^N) is deleting a link until we reach the empty network. (2) Next, starting from the empty network, we build up a sequence of networks towards g^N so that, at each step, links are only added, and players that are adding links are strictly better off at g^N compared to the current network. Then, (1) and (2) implies that from any possible network there is a farsighted improving path leading to the complete network g^N.

In the complete network we have that $Y_i(g^N) = y_i(g^N) = B/n$ because $p_i(g^N) = 0$ for all $i \in N$. Notice that expected payoffs of members of a component do not depend on how other components are structured (are linked). They only depend on the number of links and on the number of criminals within the component. That is, criminal i who belongs to group $S \in \prod(g)$ will get $Y_i(g) = y_i(g)[1 - p_i(g)\phi] = (|S|/n)\alpha_i(g)B[1 - (n - 1 - n_i)\phi/n]$, with $\alpha_i(g) = [\#\{j \in S \mid n_j = \overline{n}(S)\}]^{-1}$ if $n_i = \overline{n}(S)$ and $\alpha_i(g) = 0$ otherwise.

(1) Take any $g \neq g^N$. Case 1: For all $S \in \prod(g)$, for every $i,j \in S$, we have that $n_i = n_j$. Then, for every $i \in N$, $Y_i(g) = B/n[1 - (n - 1 - n_i)\phi/n] < B/n = Y_i(g^N)$ and thus everyone who has a link is willing to delete a link looking forward to g^N. Case 2: There exists $S \in \prod(g)$ such that $n_i < \overline{n}(S)$. Then, player i gets $Y_i(g) = 0$, so i wants to delete a link looking forward to g^N. We can repeat the arguments of Case 1 and 2 until we reach the empty network.

(2) Once we have reached the empty network we build up a sequence of networks towards g^N as follows. For $k = 1, \ldots, n$, we successively build networks so that the subset of players $\{1, \ldots, k\}$ forms a complete component, and players $k + 1, \ldots, n$ are singletons (do not have any link). We start with the empty network denoted g^1. Adding to g^1 the link $\{1,2\}$ leads to the network g^2. Notice that $\{1,2\}$ is a complete component of g^2, whereas players $3, \ldots, n$ are singletons. Let g^k, for some $k \in \{1, \ldots, n\}$, be a network such that the subset of players $\{1, \ldots, k\}$ forms a complete component, and players $k + 1, \ldots, n$ are singletons. To g^k we add successively the links $\{1, k + 1\}, \{2, k + 1\}, \ldots, \{k, k + 1\}$ to obtain the network g^{k+1}. Along this sequence of networks, the players that are adding a link are strictly better off at g^N compared to what they obtain at the current network. Indeed, when involved in adding a link, player 1 has a payoff of $B/n[1 - (n - k)\phi/n] < B/n$, player $k + 1$ has a payoff of $B/n[1 - (n - 1)\phi/n] < B/n$ (before linking to player 1) or 0 (before linking to players $\{2, ..., k\}$), and the other players have a payoff of 0.

Thus, for all $g \neq g^N$ we have $g^N \in F(g)$. Applying Theorem 4, we conclude that $\{g^N\}$ is a pairwise farsightedly stable set. ■

6. The Symmetric Connections Model and the Co-Author Model

Example 2 – *Symmetric connections model* (Jackson and Wolinsky, 1996). Players form links with each other in order to exchange information. If player i is connected to player j by a path of t links, then player i receives a payoff of δ^t from his indirect connection with player j. It is assumed that $0 < \delta < 1$, and so the payoff δ^t decreases as the path connecting players i and j increases; thus information that travels a long distance becomes diluted and is less valuable than information obtained from a closer neighbor. Each direct link ij results in a cost c to both i and j. This cost can be interpreted as the time a player must spend with another player in order to maintain a direct link. Player i's payoff from a network g is given by

$$Y_i(g) = \sum_{j \neq 1} \delta^{t(ij)} - \sum_{j\,:\,ij \in g} c,$$

where $t(ij)$ is the number of links in the shortest path between i and j (setting $t(ij) = \infty$ if there is no path between i and j).

In Figure 2 we have depicted the 3-player case where (i) for $c < \delta(1 - \delta)$, the complete network (g_7 in Figure 2) is the unique pairwise stable network and $\{g_7\}$ is the pairwise myopically stable set, (ii) for $\delta(1 - \delta) < c < \delta$, the star networks ($g_4, g_5, g_6$ in Figure 2) are pairwise stable and $\{g_4, g_5, g_6\}$ is the pairwise myopically stable set, and (iii) for $c > \delta$, the empty network is the unique pairwise stable network and $\{g_0\}$ is the pairwise myopically stable set.

Applying our concept of pairwise farsightedly stable sets to the symmetric connections model with three players, we obtain that a network g is pairwise stable if and only if $\{g\}$ is a pairwise farsightedly stable set. First we consider the case $c < \delta(1 - \delta)$. It holds that $F(g_0) = \{g_1, g_2, g_3, g_4, g_5, g_6, g_7\}$, $F(g_1) = \{g_4, g_5, g_6, g_7\}$, $F(g_2) = \{g_4, g_5, g_6, g_7\}$, $F(g_3) = \{g_4, g_5, g_6, g_7\}$, $F(g_4) = \{g_5, g_6, g_7\}$, $F(g_5) = \{g_4, g_6, g_7\}$, $F(g_6) = \{g_4, g_5, g_7\}$, and $F(g_7) = \varnothing$. Now it follows by Corollary 1 that $\{g_7\}$ is the unique pairwise farsightedly stable set.

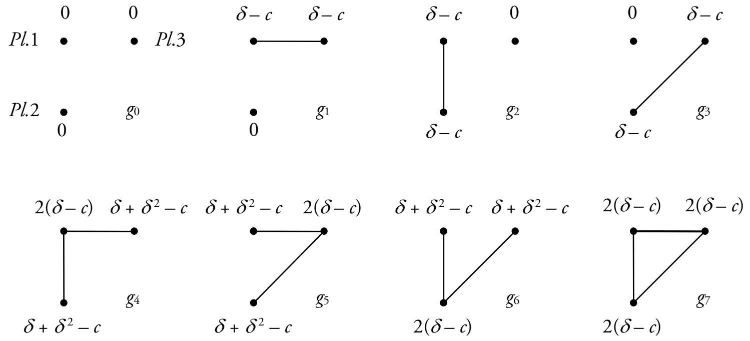


Figure 2. The symmetric connections model with three players

Next we consider the case $\delta(1-\delta) < c < \delta$. It holds that $F(g_0) = \{g_1, g_2, g_3, g_4, g_5, g_6\}$, $F(g_1) = \{g_4, g_5, g_6\}$, $F(g_2) = \{g_4, g_5, g_6\}$, $F(g_3) = \{g_4, g_5, g_6\}$, $F(g_4) = \{g_5, g_6\}$, $F(g_5) = \{g_4, g_6\}$, $F(g_6) = \{g_4, g_5\}$, and $F(g_7) = \{g_4, g_5, g_6\}$.

By a repeated application of Theorem 4, it follows that $\{g_4\}$, $\{g_5\}$, and $\{g_6\}$ are pairwise farsightedly stable sets.

Finally, we examine the case $c > \delta$. One may verify that $F(g_0) = \varnothing$, $F(g_1) = \{g_0\}$, $F(g_2) = \{g_0\}$, $F(g_3) = \{g_0\}$, $F(g_4) = \{g_0, g_1, g_2\}$, $F(g_5) = \{g_0, g_1, g_3\}$, $F(g_6) = \{g_0, g_2, g_3\}$, and $F(g_7) = \{g_0, g_1, g_2, g_3, g_4, g_5, g_6\}$. It follows by Corollary 1 that $\{g_0\}$ is the unique pairwise farsightedly stable set.

Example 3 – *Co-author model* (Jackson and Wolinsky, 1996). Each player is a researcher who spends time writing papers. If two players are connected, then they are working on a paper together. The amount of time researcher i spends on a given project is inversely related to the number of projects, n_i, that he is involved in. Formally, player i's payoff is given by

$$Y_i(g) = \sum_{j:\, ij \in g} \left(\frac{1}{n_i} + \frac{1}{n_j} + \frac{1}{n_i n_j} \right)$$

for $n_i > 0$. For $n_i = 0$ we assume that $Y_i(g) = 0$. In Figure 3 we have depicted the 3-player case. It is easily verified that the complete network g_7 is the unique pairwise stable network. Moreover, it is easy to demonstrate that the pairwise myopically stable set is $\{g_7\}$.

No singleton set is pairwise farsightedly stable in Example 3. Indeed, there is no network such that there is a farsighted improving path from any other network leading to it. More precisely, $F(g_0) = \{g_1, g_2, g_3, g_4, g_5, g_6\}$, $F(g_1) = \{g_4, g_5\}$, $F(g_2) =\{g_4, g_6\}$, $F(g_3) = \{g_5, g_6\}$, $F(g_4) = \{g_7\}$, $F(g_5) = \{g_7\}$, $F(g_6) = \{g_7\}$, and $F(g_7) = \varnothing$. However, a set formed by the complete and two star networks is a pairwise farsightedly stable set of networks. The pairwise farsightedly stable sets are $\{g_4, g_5, g_7\}$, $\{g_4, g_6, g_7\}$, $\{g_5, g_6, g_7\}$, and $\{g_1, g_2, g_3, g_7\}$ in the co-author model with three players.

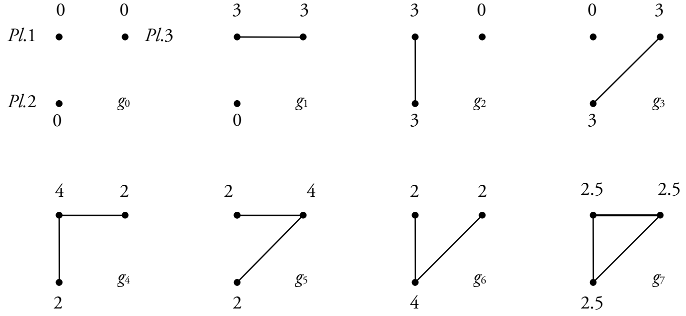


Figure 3. The co-author model with three players

7. Efficiency and Farsighted Stability

We now turn to the question of the relationship between farsighted stability and efficiency of networks. A first result is that the set of pairwise farsightedly stable networks and the set of strongly efficient networks, those which are socially optimal, may be disjoint for all allocation rules that are component balanced and anonymous.[10]

Theorem 6. *There exists a value function such that for every component balanced and anonymous rule, strongly efficient networks are not included in any of the pairwise farsightedly stable sets.*

Proof. Take the following value function defined for any $g \in \mathbb{G}$: $v(\{12, 13, 23\}) = 9$, $v(\{12, 13\}) = 0$, $v(\{12, 23\}) = 0$, $v(\{13, 23\}) = 0$, $v(\{12\}) = 8$, $v(\{13\}) = 8$, $v(\{23\}) = 8$, and $v(\varnothing) = 0$. Fix any component balanced and anonymous allocation rule Y. Then, by component balance and anonymity,

(i) $Y_1(\{12, 13, 23\}, v) = Y_2(\{12, 13, 23\}, v) = Y_3(\{12, 13, 23\}, v) = 3$,

(ii) $Y_1(\{12, 23\}, v) = c$, $Y_3(\{12, 23\}, v) = c$, $Y_2(\{12, 23\}, v) = -2c$, $Y_2(\{12, 13\}, v) = c$, $Y_3(\{12, 13\}, v) = c$, $Y_1(\{12, 13\}, v) = -2c$, $Y_1(\{13, 23\}, v) = c$, $Y_2(\{13, 23\}, v) = c$, $Y_3(\{13, 23\}, v) = -2c$,

(iii) $Y_1(\{12\}, v) = Y_2(\{12\}, v) = 4$, $Y_3(\{12\}, v) = 0$, $Y_1(\{13\}, v) = Y_3(\{13\}, v) = 4$, $Y_2(\{13\}, v) = 0$, $Y_2(\{23\}, v) = Y_3(\{23\}, v) = 4$, $Y_1(\{23\}, v) = 0$, and

(iv) $Y_1(\varnothing, v) = Y_2(\varnothing, v) = Y_3(\varnothing, v) = 0$.

The unique strongly efficient network is $\{12, 13, 23\}$. We have:

(i) $F(\varnothing) = \{\{12\}, \{13\}, \{23\}, \{12, 13, 23\}\}$;

(ii) $F(\{12\}) = \{\{13\}, \{23\}\}$, $F(\{13\}) = \{\{12\}, \{23\}\}$, $F(\{23\}) = \{\{12\}, \{13\}\}$;

(iii) For $c < 3$, $F(\{12, 13\}) = \{\{12\}, \{13\}, \{23\}, \{12, 13, 23\}\}$, for $3 \leq c < 4$, $F(\{12, 13\}) = \{\{12\}, \{13\}, \{23\}\}$, and for $c \geq 4$, $F(\{12, 13\}) = \{\{12\}, \{13\}\}$. Next, for $c < 3$, $F(\{12, 23\}) = \{\{12\}, \{13\}, \{23\}, \{12, 13, 23\}\}$, for $3 \leq c <4$, $F(\{12, 23\}) = \{\{12\}, \{13\}, \{23\}\}$, and for $c \geq 4$, $F(\{12, 23\}) = \{\{12\}, \{23\}\}$. And, for $c < 3$, $F(\{13, 23\}) = \{\{12\}, \{13\}, \{23\}, \{12, 13, 23\}\}$, for $3 \leq c < 4$, $F(\{13, 23\}) = \{\{12\}, \{13\}, \{23\}\}$, and for $c \geq 4$, $F(\{13, 23\}) = \{\{13\}, \{23\}\}$;

(iv) For $c < 3$, $F(\{12, 13, 23\}) = \{\{12\}, \{13\}, \{23\}, \{12, 13\}, \{12, 23\}, \{13, 23\}\}$, for $c \geq 3$, $F(\{12, 13, 23\}) = \{\{12\}, \{13\}, \{23\}\}$.

Thus, $\{\{12\}\}$, $\{\{13\}\}$, and $\{\{23\}\}$ are the only pairwise farsightedly stable sets. ■

[10] Bhattacharya (2005) has obtained a similar result with respect to the notion of the largest consistent set.

A second result considers the case where there is a network that strictly Pareto dominates all other networks. That is, if there is a network g such that for all $g' \in \mathbb{G} \setminus \{g\}$ it holds that, for all i, $Y_i(g, v) > Y_i(g', v)$. Although the network that strictly Pareto dominates all others is pairwise stable, there might be many more pairwise stable networks. We will show in Section 8 that also the concept of the largest pairwise consistent set suffers from a similar defect. The following result asserts that pairwise farsighted stability singles out the Pareto dominating network as the unique pairwise farsightedly stable set.

Theorem 7. *If there is a network g that strictly Pareto dominates all other networks, then* $\{g\}$ *is the unique pairwise farsightedly stable set.*

Proof. It is immediate that $g \in F(g')$ for all $g' \in \mathbb{G} \setminus \{g\}$ and that $F(g) = \emptyset$. Corollary 1 leads to the desired result. ■

We next provide sufficient conditions on the allocation rule and/or the value function such that there is no conflict between strong efficiency and farsighted stability.

An immediate application of Theorem 7 is the case of increasing returns to link creation as defined in Dutta et al. (2005). This property requests that along every nested sequence of increasingly connected networks, there is a threshold network for which the value turns nonnegative, and both aggregate as well as payoffs of individuals who form extra links then increase as the network becomes even larger. Under this condition, and with a componentwise egalitarian allocation rule, g^N Pareto dominates all other networks, so Theorem 7 applies.

An allocation rule is said to be egalitarian if for every $v \in \mathcal{V}$ and $g \in \mathbb{G}$, $Y_i(g, v) = v(g)/n$. The following result follows as a corollary to Theorem 7.

Corollary 2. *Suppose that Y is the egalitarian rule and there is a unique strongly efficient network* g^e. *Then,* $\{g^e\}$ *is the unique pairwise farsightedly stable set.*

8. The von Neumann-Morgenstern Pairwise Farsightedly Stable Set

The pairwise farsightedly stable set requires deterrence of external deviations, external stability, and minimality. The von Neumann-Morgenstern stable set (von Neumann and Morgenstern, 1953) imposes internal and external stability. Incorporating the notion of farsighted improving paths into the original definition of the von Neumann-Morgenstern stable set, we obtain the von Neumann-Morgenstern pairwise farsightedly stable set.

Definition 5. The set G is a von Neumann-Morgenstern pairwise farsightedly stable set if (i) $\forall g \in G$, $F(g) \cap G = \emptyset$ and (ii) $\forall g' \in \mathbb{G} \setminus G$, $F(g') \cap G \neq \emptyset$.

Von Neumann-Morgenstern pairwise farsightedly stable sets do not always exist. Consider the situation where three players can form links and where the payoffs are given in Figure 4. We have $F(g_0) = F(g_7) = \{g_1, g_2, g_3, g_4, g_5, g_6\}$, $F(g_1) = \{g_2, g_3\}$, $F(g_2) = \{g_3, g_4, g_5\}$, $F(g_3) = \{g_4, g_5\}$, $F(g_4) = \{g_1, g_5, g_6\}$, $F(g_5) = \{g_1, g_6\}$, $F(g_6) = \{g_1, g_2, g_3\}$. We prove that there is no von Neumann-Morgenstern pairwise farsightedly stable set. Suppose on the contrary that G is a von Neumann-Morgenstern pairwise farsightedly stable set. Suppose $g_1 \in G$. The internal stability condition implies that no other network can belong to G. Since $F(g_2) \cap \{g_1\} = \varnothing$ it follows that external stability is violated, a contradiction. As a consequence, $g_1 \notin G$. A symmetric argument leads to the result that $g_3 \notin G$ and $g_5 \notin G$. Suppose now that $g_2 \in G$. The internal stability condition implies that no other network can belong to G. Since $F(g_5) \cap \{g_2\} = \varnothing$ it follows that external stability is violated, a contradiction. By symmetry it follows that $g_4 \notin G$ and $g_6 \notin G$. Suppose $g_0 \in G$. Internal stability implies that no other network can belong to G. Since $F(g_1) \cap \{g_0\} = \varnothing$, it follows that external stability is violated, a contradiction. By a similar argument we can show that $g_7 \notin G$. The only remaining possibility is $G = \varnothing$. This clearly violates external stability. Thus, there is no von Neumann-Morgenstern pairwise farsightedly stable set in this example. However, pairwise farsightedly stable sets do exist by virtue of Theorem 3. The pairwise farsightedly stable sets are $\{g_1, g_2, g_3\}$, $\{g_3, g_4, g_5\}$, and $\{g_1, g_5, g_6\}$.

From Theorem 3 we immediately obtain the following corollary, which states that a von Neumann-Morgenstern pairwise farsightedly stable set is also a pairwise farsightedly stable set.

Corollary 3. *If G is a von Neumann-Morgenstern pairwise farsightedly stable set, then G is a pairwise farsightedly stable set.*

Since internal stability is automatically satisfied when a set of networks contains only one element, Theorem 4 leads to the following corollary.

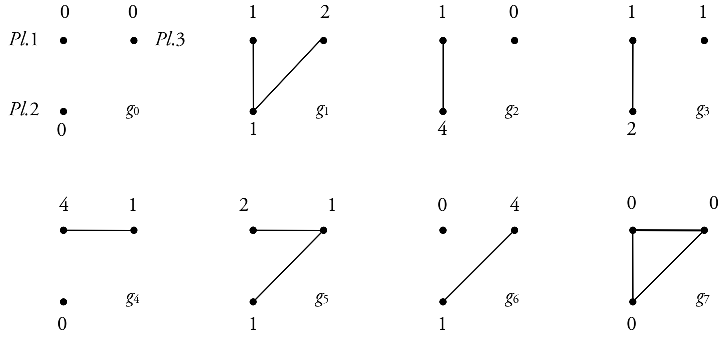


Figure 4. Non-existence of von Neumann-Morgenstern pairwise farsightedly stable sets

Corollary 4. *The set* $\{g\}$ *is a pairwise farsightedly stable set if and only if it is a von Neumann-Morgenstern pairwise farsightedly stable set.*

From Theorem 5 and Corollary 3 we immediately get the next result, which implies the converse of Corollary 3, a unique pairwise farsightedly stable set is also a von Neumann-Morgenstern pairwise farsightedly stable set.

Corollary 5. *If* G *is the unique pairwise farsightedly stable set, then* G *is the unique von Neumann-Morgenstern pairwise farsightedly stable set.*

We have shown that replacing the internal stability condition in the von Neumann-Morgenstern pairwise farsightedly stable set by deterrence of external deviations and minimality, leads to a stability concept that contains the von Neumann-Morgenstern pairwise farsightedly stable set, and is always non-empty.

EXAMPLE 1 – *Criminal networks*, $n = 3$ (*continued*). Since $\{g_7\}$ is the unique pairwise farsightedly stable set, Corollary 4 shows that $\{g_7\}$ is the unique von Neumann-Morgenstern pairwise farsightedly stable set.

EXAMPLE 3 – *Co-author model (continued).* By Corollary 3 we have to analyze whether any of the pairwise farsightedly stable sets is a von Neumann-Morgenstern pairwise farsightedly stable set. The unique von Neumann-Morgenstern pairwise farsightedly stable set is given by $\{g_1, g_2, g_3, g_7\}$. The other pairwise farsightedly stable sets $\{g_4, g_5, g_7\}$, $\{g_4, g_6, g_7\}$, $\{g_5, g_6, g_7\}$, are not von Neumann-Morgenstern pairwise farsightedly stable sets because they do not satisfy the internal stability condition.

9. The Largest Pairwise Consistent Set

In this section we study the relationship between pairwise farsighted stability and the largest pairwise consistent set, a concept that has been defined in Chwe (1994) for general social environments. By considering a network as a social environment, and by allowing only pairwise deviations, we obtain the definition of the largest pairwise consistent set.

Definition 6. G is a pairwise consistent set if $\forall g \in G$,

(ia) $\forall ij \notin g$, $\exists g' \in G$, where $g' = g + ij$ or $g' \in F(g + ij) \cap G$, such that $Y_i(g', v) < Y_i(g, v)$ or $Y_j(g', v) < Y_j(g, v)$ or $(Y_i(g', v), Y_j(g', v)) = (Y_i(g, v), Y_j(g, v))$,

(ib) $\forall ij \in g$, $\exists g', g'' \in G$, where $g' = g - ij$ or $g' \in F(g - ij) \cap G$, and $g'' = g - ij$ or $g'' \in F(g - ij) \cap G$, such that $Y_i(g', v) \leq Y_i(g, v)$ and $Y_j(g'', v) \leq Y_j(g, v)$.

The largest pairwise consistent set is the pairwise consistent set that contains any pairwise consistent set.

The set G is a pairwise consistent set if both external and internal deviations are deterred. The largest pairwise consistent set is the set that contains any pairwise consistent set. It follows from the results in Chwe (1994) that the largest pairwise consistent set exists, is non-empty, and satisfies external stability.

Our pairwise farsightedly stable sets need not be consistent in the sense of Chwe (1994) since we do not require internal deviations to be deterred. Moreover a pairwise consistent set does not necessarily satisfies the external stability condition. Only the largest pairwise consistent set is guaranteed to satisfy external stability.

Whenever G is a von Neumann-Morgenstern pairwise farsightedly stable set, G is also a pairwise farsightedly stable set (Corollary 3) and a largest pairwise consistent set (see Chwe, 1994). Replacing the internal and external stability conditions of the von Neumann-Morgenstern pairwise farsightedly stable set by the conditions that internal and external deviations should be deterred, Chwe (1994) has proposed a stability concept that always exists and that contains the von Neumann-Morgenstern pairwise farsightedly stable set. In this paper, replacing the internal stability condition by the condition that external deviations should be deterred and the minimality condition, we propose another stability concept that also contains the von Neumann-Morgenstern pairwise farsightedly stable set.

Chwe (1994) provides the following iterative procedure to find the largest consistent set. Let $Z^0 \equiv \mathbb{G}$. Then, Z^k ($k = 1,2,\ldots$) is inductively defined as follows: $g \in Z^{k-1}$ belongs to Z^k with respect to Y and v if

(ia) $\forall ij \notin g$, $\exists g' \in Z^{k-1}$, where $g' = g + ij$ or $g' \in F(g + ij)$ such that $Y_i(g', v) < Y_i(g, v)$ or $Y_j(g', v) < Y_j(g, v)$ or $(Y_i(g', v), Y_j(g', v)) = (Y_i(g, v), Y_j(g, v))$.

(ib) $\forall ij \in g$, $\exists g', g'' \in Z^{k-1}$, where $g' = g - ij$ or $g' \in F(g - ij)$, and $g'' = g - ij$ or $g'' \in F(g - ij)$, such that $Y_i(g', v) \leq Y_i(g, v)$ and $Y_j(g'', v) \leq Y_j(g, v)$.

The largest pairwise consistent set is given by $\bigcap_{k \geq 1} Z^k$. That is, a network $g \in Z^{k-1}$ is stable (at step k) and belongs to Z^k, if all possible pairwise deviations are deterred. Consider a pairwise deviation from g that involves making the link ij. There might be further pairwise deviations which end up at g', where $g + ij \to g'$. If either i or j is worse off at g' or both are equally well off compared to the original network g, then the pairwise deviation is deterred. Similarly for a pairwise deviation from g that involves deleting the link ij. There might be further pairwise deviations which end up at g' and g'' where $g - ij \to g'$ and $g - ij \to g''$. If i is equally well or worse off at g' and j is equally well or worse off at g'' compared to the original network g, then the pairwise deviation is deterred. Since $\mathbb{G}$ is finite, there exists $m \in N$ such that $Z^k = Z^m$ for all $k \geq m$, and Z^m is the largest pairwise consistent set.

The next result states that if a network is not in the largest pairwise consistent set, it cannot be a pairwise farsightedly stable set of networks.

Theorem 8. *If $\{g\}$ is a pairwise farsightedly stable set, then g belongs to the largest pairwise consistent set.*

Proof. Since $\{g\}$ is a pairwise farsightedly stable set we have that for all $ij \notin g$: $g \in F(g + ij)$ and for all $ij \in g$: $g \in F(g - ij)$. So $g \in Z^1$. By induction, $g \in Z^k$ for $k \geq 1$. So, g belongs to the largest pairwise consistent set. ■

Remember that two networks g and g' are adjacent if they differ by one link. The value function v and allocation rule Y exhibit no *indifference* if for any g and g' that are adjacent either g defeats g' or g' defeats g.

Theorem 9. *Suppose that Y and v exhibit no indifference. If g is pairwise stable then it belongs to the largest pairwise consistent set.*

Proof. Since Y and v exhibit no indifference, we have that a pairwise stable network g defeats (i) $g + ij$ for all $ij \notin g$ and (ii) $g - ij$ for all $ij \in g$. Thus, $g \in F(g + ij)$ and $g \in F(g - ij)$. So $g \in Z^1$. By induction $g \in Z^k$ for $k \geq 1$. So, g belongs to the largest pairwise consistent set. ■

We claimed in Section 7 that even if there is a network that strictly Pareto dominates all other networks, the largest pairwise consistent set may contain other networks. It is not difficult to construct examples where the no indifference property holds, and some network strictly Pareto dominates all others. Moreover, such an example can be constructed such that inefficient networks are pairwise stable. It then follows from Theorem 9 that such a network also belongs to the largest pairwise consistent set. By virtue of Theorem 7, such a network does not belong to any pairwise farsightedly stable set.

Let us calculate the largest pairwise consistent set in two examples.

EXAMPLE 1 – *Criminal networks*, $n = 3$, $\phi < 3/2$ (*continued*). We apply the iterative procedure of Chwe (1994) to find the largest pairwise consistent set. We start with $Z^0 = \{g_0, g_1, \ldots, g_7\}$.

Starting in g_0, players 1 and 2 can add the link {1,2} to move to g_2. The indirect dominance relation implies that from there it is only possible to end up in g_1, g_3, or g_7. In all these networks both players have at least the same payoffs as in g_0, and at least one player has strictly higher payoffs. It follows that $g_0 \in Z^1$. Any deviation from g_1 may lead back to g_1 and is thereby deterred, so $g_1 \in Z^1$. By symmetry it holds that $g_2, g_3 \in Z^1$. Starting in g_4, players 2 and 3 can add the link

{2, 3}, resulting in g_7. Since $F(g_7) = \varnothing$ and both players have higher payoffs in g_7 than in g_4, we find that $g_4 \notin Z^1$. By symmetry it holds that g_5, $g_6 \notin Z^1$. Any deviation from g_7 may lead back to g_7 and is thereby deterred, so $g_7 \in Z^1$. The same arguments can be used to show that $Z^2 = Z^1$. It follows that the largest pairwise consistent set equals $\{g_1, g_2, g_3, g_7\}$.

Theorem 8 implies that g_7 belongs to the largest pairwise consistent set. This example therefore demonstrates that the largest pairwise consistent set may contain other networks too.

EXAMPLE 3 – *Co-author model* (*continued*). We apply the iterative procedure of Chwe (1994) to find the largest pairwise consistent set. We start with $Z^0 = \{g_0, g_1, \ldots, g_7\}$. Starting in g_0, players 1 and 2 can add the link {1,2} to move to g_2. The indirect dominance relation implies that from there it is only possible to reach g_4 or g_6. In all these networks, players 1 and 2 have higher payoffs than at g_0. It follows that $g_0 \notin Z^1$. Starting in g_4, players 2 and 3 will add a link to move to g_7. Since $F(g_7) = \varnothing$, no further moves will occur. Players 2 and 3 have higher payoffs at g_7 than at g_4. It follows that $g_4 \notin Z^1$. For similar reasons, $g_5 \notin Z^1$ and $g_6 \notin Z^1$. It can be verified that $Z^1 = \{g_1, g_2, g_3, g_7\}$.

We show next that $Z^2 = \{g_1, g_2, g_3, g_7\}$. Starting in g_1, players 1 and 2 may add a link to go to g_4, a network not in Z^1. From g_4 the indirect dominance relation dictates a move to g_7. In g_7 player 1 is worse off than in g_1. It follows that no link will be added by them to g_1. Repeating such arguments, it can be shown that $Z^2 = \{g_1, g_2, g_3, g_7\} = Z^k$ for all $k \geq 2$. It follows that the largest pairwise consistent set equals $\{g_1, g_2, g_3, g_7\}$.

Since the assumptions of Theorem 9 are satisfied in this example, it follows that g_7 belongs to the largest pairwise consistent set. The example shows that the largest pairwise consistent set may contain other networks too.

Table 1 summarizes our findings in Example 1 with $n = 3$ and $\phi < 3/2$ and Example 3.

Table 1. The (no)-relationships among solution concepts for network stability

Concept	Example 1	Example 3
Pairwise myopically stable set	$\{g_1, g_2, g_3, g_7\}$	$\{g_7\}$
Pairwise farsightedly stable set	$\{g_7\}$	$\{g_4, g_5, g_7\}, \{g_4, g_6, g_7\}, \{g_5, g_6, g_7\}, \{g_1, g_2, g_3, g_7\}$
vN-M farsighted stable set	$\{g_7\}$	$\{g_1, g_2, g_3, g_7\}$
Largest pairwise consistent set	$\{g_1, g_2, g_3, g_7\}$	$\{g_1, g_2, g_3, g_7\}$

10. Conclusion

We have proposed a new concept, the pairwise farsightedly stable set, to predict which networks may be formed among farsighted players. A set of networks G is pairwise farsightedly stable (i) if all possible pairwise deviations from any network $g \in G$ to a network outside G are deterred by the threat of ending worse off or equally well off, (ii) if there exists a farsighted improving path from any network outside the set leading to some network in the set, and (iii) if there is no proper subset of G satisfying conditions (i) and (ii). We have shown that a pairwise farsightedly stable set always exists and we provide a full characterization of unique pairwise farsightedly stable sets of networks. As a corollary we have given the necessary and sufficient condition such that a unique pairwise farsightedly stable set consisting of a single network exists. We have found that the pairwise farsightedly stable sets and the set of strongly efficient networks may be disjoint. Nevertheless, contrary to other pairwise concepts, if there is a network that Pareto dominates all other networks, then that network is the unique prediction of pairwise farsighted stability. We have also been able to provide some conditions on the allocation rule and the value function such that pairwise farsighted stability singles out the strongly efficient network. Finally, we have studied the relationship between pairwise farsighted stability and other concepts such as pairwise myopic stability, the von Neumann-Morgenstern pairwise farsightedly stable set, and the largest pairwise consistent set. When there is a unique pairwise farsightedly stable set, then its elements are also pairwise stable. Pairwise stable networks belong to the largest pairwise consistent set when a mild no indifference criterion is satisfied. Moreover, any von Neumann-Morgenstern pairwise farsightedly stable set is also a pairwise farsightedly stable set. A pairwise farsightedly stable set consisting of a unique element is also a von Neumann-Morgenstern pairwise farsightedly stable set. If there is a unique pairwise farsightedly stable set, then it is also the unique von Neumann-Morgenstern pairwise farsightedly stable set. By means of examples we have shown that there is no general relationship between (i) pairwise farsightedly stable sets and pairwise myopically stable sets and (ii) pairwise farsightedly stable sets and largest pairwise consistent sets.

References

Aumann, R. and R. Myerson (1988), 'Endogenous formation of links between players and coalitions: An application of the Shapley value', in Roth, A. (ed.), *The Shapley Value*, Cambridge: Cambridge University Press, pp. 175–191.

Bhattacharya, A. (2005), 'Stable and efficient networks with farsighted players: The largest consistent set', Mimeo, University of York.

Calvó-Armengol, A. and Y. Zenou (2004), 'Social networks and crime decisions: The role of social structure in facilitating delinquent behavior', *International Economic Review* **45**, 939–958.

Chwe, M.S. (1994), 'Farsighted coalitional stability', *Journal of Economic Theory* **63**, 299–325.

Dutta, B., S. Ghosal and D. Ray (2005), 'Farsighted network formation', *Journal of Economic Theory* **122**, 143–164.

Dutta, B. and S. Mutuswami (1997), 'Stable networks', *Journal of Economic Theory* **76**, 322–344.

Herings, P.J.J., A. Mauleon and V. Vannetelbosch (2004), 'Rationalizability for social environments', *Games and Economic Behavior* **49**, 135–156.

Jackson, M.O. (2003), 'The stability and efficiency of economic and social networks', in B. Dutta and M.O. Jackson (eds.), *Networks and Groups: Models of Strategic Formation*, Heidelberg: Springer-Verlag, pp. 99–140.

Jackson, M.O. (2005), 'A survey of models of network formation: Stability and efficiency', in G. Demange and M. Wooders (eds.), *Group Formation in Economics: Networks, Clubs and Coalitions*, Cambridge: Cambridge University Press, pp. 11–57.

Jackson, M.O. and A. van den Nouweland (2005), 'Strongly stable networks', *Games and Economic Behavior* **51**, 420–444.

Jackson, M.O. and A. Watts (2002), 'The evolution of social and economic networks', *Journal of Economic Theory* **106**, 265–295.

Jackson, M.O. and A. Wolinsky (1996), 'A strategic model of social and economic networks', *Journal of Economic Theory* **71**, 44–74.

Mauleon, A. and V. Vannetelbosch (2004), 'Farsightedness and cautiousness in coalition formation games with positive spillovers', *Theory and Decision* **56**, 291–324.

Page Jr., F.H., M. Wooders and S. Kamat (2005), 'Networks and farsighted stability', *Journal of Economic Theory* **120**, 257–269.

Page Jr., F.H. and M. Wooders (2005), 'Strategic basins of attraction, the path dominance core, and network formation games', Working Paper 05-W09, Department of Economics, Vanderbilt University.

Sengupta, A. and K. Sengupta (1994), 'Viable proposals', *International Economic Review* **35**, 347–359.

Tercieux, O. and V. Vannetelbosch (2006), 'A characterization of stochastically stable networks', *Int. J. Game Theory* **34**, 351–369.

von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Watts, A. (2002), 'Non-myopic formation of circle networks', *Economics Letters* **74**, 277–282.

Xue, L. (1998), 'Coalitional stability under perfect foresight', *Economic Theory* **11**, 603–627.

Efficiency versus Stability in Climate Coalitions: A Conceptual and Computational Appraisal

Thierry Bréchet, François Gerard and Henry Tulkens

Paper presented at CTN Workshop 2010. Reprinted from *The Energy Journal* **32**(1), 2011, 49–76, with permission from IAEE.

This paper evaluates with numerical computations the respective merits of two competing notions of coalition stability in the standard global public goods model of climate change. To this effect it uses the CWS integrated assessment model. After a reminder of the two game theoretical stability notions involved – core-stability and internal-external stability – and of the CWS model, the former property is shown to hold for the grand coalition if resource transfers of a specific form between countries are introduced. The latter property appears to hold neither for the grand coalition nor for most large coalitions whereas it is verified for most small coalitions in a weak sense that involves transfers. Finally, coalitions, stable in either sense, that perform best in terms of carbon concentration and global welfare are always heterogeneous ones. Therefore, if coalitional stability is taken as an objective, promoting small or homogeneous coalitions is not to be recommended.

1. Introduction

The global public good character of combating the effects of climate change requires voluntary cooperation amongst countries if any improvement upon the *laissez faire* business-as-usual is sought for. Such cooperation, institutionalized in international environmental treaties, consists in joint actions decided and implemented by the signatory countries. Negotiated under the United Nations Framework Convention on Climate Change (UNFCCC), the Kyoto Protocol represents the first legally binding agreement on climate. As such, it is now considered as a decisive step. However it

The authors wish to thank Johan Eyckmans for the collaboration under the CLIMNEG project. This paper is part of the CLIMNEG project supported by the Belgian Science Policy (contract SD/CP/05A).

is widely acknowledged that, in order to be environmentally effective, post-Kyoto agreements should include more countries and yield stronger carbon emission abatement. This twin issue (which countries and how much more abatement?) is at the heart of the on-going negotiation process that currently prepares, under the UNFCCC, for the post 2012 world climate regime.

Calling a 'coalition' any set of countries thus joining their efforts against climate change, an abundant literature has developed over the last 15 years dealing with the issue of the likeliness of 'stable' climate coalitions. In that literature, two stability concepts are competing: the core-stability and the internal-external stability.[1] An early summary of that competition was reported in Tulkens (1998) with an update in Chander and Tulkens (2009). In brief, the core-stability concept focuses on strategies chosen by the members of the grand coalition, which gather all countries. By contrast, the internal-external stability focuses on strategies chosen by any coalitions of any size, and evaluates the benefits for each country of being inside or outside these coalitions.[2] Formal definitions are provided in Section 2.2 below. Up to now, the confrontation of the two concepts has been exclusively in terms of their logical properties.

In this paper we wish to make the confrontation at the level of an application, and discuss some policy implications. For that purpose we make use of a dynamic numerical integrated assessment model, namely the ClimNeg World Simulation (henceforth CWS) model, which lends itself to proceed fairly easily to the comparison we are interested in. Such a numerical approach of the coalitional stability problem has been initiated in Eyckmans and Tulkens (2003), who actually introduced the CWS model and used it to explore one of the two conceptual approaches just mentioned. This was followed and pursued in Carraro et al. (2006), who explored with CWS the other approach.[3] By putting together these two explorations with an updated version of the CWS model, the present paper presents an explicit comparison, with the purpose of bringing to light the properties of potential coalitions in three respects: stability, climate performance and global welfare.

The contribution of our paper is twofold. First, it is methodological. By testing on the same integrated assessment model the two alternative game theoretic stability concepts, we better show their respective merits, most typically in terms of existence of stable coalitions in either sense. Second, the paper contributes to the policy debate. Assessing the properties of alternative climate coalitions in a concrete numerical context gives a powerful justification for recommendations as

[1] One of the two concepts is often assimilated with 'self enforcement' (of treaties signed by members of stable coalitions), as suggested initially by Barret (1994) and elaborated upon in Barret (2003). Actually, this attractive expression applies equally well to both stability concepts. There is thus no gain in using it here.

[2] In the literature it is sometimes referred to the latter as the *cooperative approach* and to the former as the *non-cooperative approach*, see e.g. Bréchet and Eyckmans (2010).

[3] There exist some other works that also use game theory, e.g. Bernard et. al. (2008) or Yang (2008).

to the size and nature (versus heterogeneity) of possible climate coalitions. Moreover, by showing explicitly which transfers among countries are appropriate to stabilize efficient coalitions, the paper also identifies a way of widening the scope of negotiations that the success of the Montreal Protocol has confirmed.

The paper is organized as follows. After this introduction, Section 2 presents the reader with the basic game theoretic concepts of coalition stability that we wish to put to a test. Section 3 presents the CWS integrated assessment model, including its calibration. Section 4 contains the main numerical results on the two alternative stability concepts when applied to the CWS model, and Sections 5 and 6 comment on the issues of homogeneity vs heterogeneity, aggregate welfare and environmental performance of alternative coalitions. Some sensitivity analyzes presented in Section 7 show the robustness of our results and the concluding Section 8 summarizes our main findings and derives their policy implications.

2. The Conceptual Framework

2.1 The climate-economic model and its associated games

The methodology we are using requires to make precise the relationship between the climate-economic model (CWS) and the games to which the alternative stability concepts are applied. In this section we deal with the game theoretic concepts while the economic model will be described in Section 3.

Two categories of games are involved, namely cooperative and non-cooperative ones. In either case the players are the countries, each player's strategies are the values chosen for the economic decision variables and the players payoffs are the countries' welfare level at the end of that period. A family of n such strategies, one for each player, defines what we call in the following section a *scenario*. Among the many conceivable ones we shall deal with (i) the Nash equilibrium scenario, (ii) various scenarios of partial agreement Nash equilibrium with respect to given coalitions, and (iii) the Pareto efficient scenario.

Non-cooperative games are those that consider strategies enacted by individual players; they lead essentially to the Nash equilibrium concept. Cooperative games, by contrast, typically consider in addition the strategies chosen jointly by groups of players, usually called coalitions, that is, subsets of players (including singletons and the all players set). In either case the behavioral assumption is made that the strategy chosen by individual players as well as the strategies chosen jointly by coalitions result from payoff maximization over some feasible set: the individual payoffs in the non-cooperative setting, the joint payoffs of the coalition members in the cooperative setting, this joint payoff being called the worth of the coalition.[4]

[4] We deal only with transferable utility (TU) games, for two reasons. On the one hand, at the theoretical

2.2 The stability concepts

The two approaches of the stability of a coalition rest on different views when applied to international environmental agreements. The core-stability approach assumes that, if one or several countries attempt to free-ride on an efficient agreement with transfers, the other countries do not cooperate among themselves anymore, so as to make the free rider(s) see that their country is better off by not free riding. This threat is what induces stability. In the internal-external stability approach, stability of an agreement within a coalition obtains if no individual country attempts to free ride on it, assuming that free riding does not prevent the other countries from keeping cooperation among themselves.

2.2.1 'Gamma core' stability

The core-stability theory focuses on strategies chosen jointly by the members of the grand coalition, that is, the set N of all players. The behavioral assumption mentioned above implies that, in the CWS model, N chooses the Pareto efficient scenario.

This scenario and the grand coalition that generates it are then said to be *stable in the core sense* if the scenario belongs to the core of a suitably defined cooperative game, that is, if it is such that (i) no individual player can reach a higher payoff by *not* adopting the strategy assigned to him in the efficient scenario and choosing instead the best individual strategy he could find; and (ii) no subset of players, smaller than N, can similarly do better for its members, that is, by rejecting the strategies assigned to them by the efficient scenario and adopting a strategy of their own. Consequently, the grand coalition N is called strategically stable and its scenario may rightly be called self *enforceable* since no coalition can find a better one for its members.

Formally, let i refer to players ($i = 1, \ldots, n$), W_i denote the payoff of player $i.$ $S \subseteq N$ denote a coalition, the scalar $W(S)$ be the worth of coalition S and the vector $W = (W_1, \ldots, W_i, \ldots, W_n)$ denote an imputation.[5] The imputation W will be said to belong to the core of the cooperative game if the individual payoffs W_i satisfy the following property:

- ***Property CR: Coalitional rationality*** $\forall S \subseteq N,\ \sum_{i \in S} W_i \geq W(S)$

Notice that this property implies:

- ***Property IR: Individual rationality*** $\forall i \in N,\ W_i \geq W(\{i\})$

level, the stability concepts we use have been developed for such games only; on the other hand, only TU games are used in applied numerical works such as this one.

[5] An imputation is any vector of individual payoffs W_i such that their sum is equal to the worth of the grand coalition, formally: $\sum_{i \in N} W_i = W(N)$. By construction it is induced by an efficient strategy.

To be complete, the formal statement of these two properties should further specify what are the players' strategies implicit in the right hand sides of these expressions, namely $W(\{i\})$ and $W(S)$. In the former, the strategy and the ensuing payoff of player i are those of the Nash equilibrium scenario; in the latter, the worth of coalition S is the sum of the payoffs obtained by the members of S as they result from enacting the joint strategy that maximizes this sum; this is the scenario dubbed above partial agreement Nash equilibrium (*PANE*) with respect to a coalition.[6]

2.2.2 Internal-external stability

Rather than focusing on strategies of the grand coalition, the internal-external stability theory considers *any* coalition S and the payoffs of its members at the corresponding *PANE* scenario.[7] It then considers the strategies and the resulting individual payoffs that can be reached by every player along that scenario according to whether he is inside or outside of the coalition.[8] Being inside means for the player to follow the strategy he is assigned to within the coalition he is a member of, whereas being outside means behaving as a singleton, taking as given the behavior of the coalition he is not a member of as well as of the other players (assumed to behave as singletons too). A coalition S and the *PANE* scenario it generates are then said to be *stable in the internal-external sense* if the scenario is such that no insider prefers to stay out of the coalition and no outsider prefers to join the coalition rather than stay aside. Consequently, the coalition S is called stable and its *PANE* scenario *self enforceable*, not by reference to alternative coalitions as in the preceding concept, but instead because of the structure of the individual motivations of the players within and outside the coalition.

Formally, letting $W_i(S)$ denote the individual payoff of player i when coalition S is formed, this means that the payoffs satisfy the following two properties:[9]

- ***IS Property (Internal Stability)***: $\forall i \in S,\ W_i(S) \geq W_i(S\backslash\{i\})$
- ***ES Property (External Stability)***: $\forall i \notin S,\ W_i(S) \geq W_i(S \cup \{i\})$

[6] In a partial agreement Nash equilibrium with respect to a coalition, the coalition members are assumed, as usual, to maximize their joint payoffs; but it is assumed in addition – and this is not usual – that the players outside of the coalition choose, as singletons, the strategy that maximizes their individual payoff, given what the coalition and the other singletons do. The equilibrium concept derived from this assumption (called the 'gamma' assumption) was introduced in Chander and Tulkens (1995) and (1997) as the essential building block of the 'gamma core' concept they proposed, which is to be used hereafter. A powerful further justification of the assumption is provided in Chander (2008).

[7] Thus, the gamma assumption is used here too.

[8] It is assumed that a player can only either join the coalition or remain alone.

[9] The internal-external stability concept originates in the work of D'Aspremont et al. (1983) and (1986) on the stability of cartels and has been imported in the literature on IEAs by Carraro and Siniscalco (1993) and Barrett (1994). The way it is presented here – in particular its connection with the PANE concept – owes much to Eyckmans and Finus (2004).

2.3 Transfer schemes

It has often been suggested that when a coalition and its strategies are not stable, transfers of payoffs (of economic goods, in economic games) between players may induce stability. To what extent is this the case for each of the two forms of stability just defined?

In the context of the core-stability theory, transfers were proposed by Chander and Tulkens (1995, 1997) for the standard game with multilateral externalities used to deal with international environmental agreements. They proved analytically that transfers formulated as follows induce the stability property.

Let W_i^{Nash} be the payoff of player i at the Nash equilibrium of the non-cooperative game, that is, in absence of cooperation; and let

$$W^*(N) = (W_1^*, \ldots, W_n^*),$$

be the payoff vector of the players at the Pareto efficient solution of the cooperative game. The transfers consist of the following payoff amounts (positive if received, negative if paid by i):

$$\Psi_i = -(W_i^* - W_i^{Nash}) + \pi_i\left(\sum_{j\in N} W_j^* - \sum_{j\in N} W_j^{Nash}\right) \quad i = 1, \ldots, n, \tag{1}$$

with $\pi_i \geq 0$ $\forall i$ such that $\sum_i \pi_i = 1$.

These transfers guarantee that each player receives a payoff at least equal to what it is in case of no cooperation and it divides the surplus of cooperation over non-cooperation according to weights π_i. In the multilateral environmental model, each weight is equal to the ratio of player i's marginal damage cost over the sum over all players of such marginal damage costs. With these weights, the payoff vector,[10] given by

$$W^*(N) + \Psi_n =_{def} (W_1^* + \Psi_1, \ldots, W_n^* + \Psi_n),$$

is shown by Chander and Tulkens (1995, 1997) to belong to the core of the game.

The internal-external stability theory proposes no specific transfer formula but introduces instead, in Eyckmans and Finus (2004), the notion of potentially internally stable coalitions. A coalition (of any size) is *potentially internally stable* if it can guarantee to all its members at least their free-rider payoff. For a given a coalition, the free-rider payoff of any of its members is the payoff the member would obtain in the *PANE* scenario *w.r.t.* that coalition if he would stay out and behave as a singleton in the face of that coalition.

Formally, for any coalition S, this reads as follows:

[10] That $W^*(N) + \Psi_n$ is an imputation follows from the fact that (1) implies $\sum_{i\in N}\Psi_i = 0$, i.e. the transfers budget balances.

- *PIS Property (Potential Internal Stability)*: $W(S) \geq \sum_{i \in S} W_i(S \backslash \{i\})$

The free rider payoff of a player *i vis-à-vis* some coalition S – that is, each term of the sum in the right hand side of the equation – may be seen as the minimum payoff player i requires to remain a member of the coalition. Coalitions whose worth under their *PANE* is large enough to meet this requirement for all their members can thus be stabilized at least internally.[11]

3. The ClimNeg World Simulation model (CWS)

3.1 Overview of the model

The ClimNeg World Simulation model (CWS) is a dynamic integrated assessment model of climate change and optimal growth, adapted for coalitional analysis from Nordhaus and Yang (1996). It encompasses economic, climatic and impact dimensions in a worldwide intertemporal setting. As a Ramsey-type model, growth is driven by population growth, technological change and capital accumulation. The time dimension is discrete, indexed by *t*, finite, but very long. The world is split into six countries/regions: USA, Japan, Europe,[12] China, the Former Soviet Union and the Rest of the World.[13] In each country/region[14] $i = 1, ..., n$ gross output is given by a Cobb-Douglas production function combining capital and population. Population is exogenous. Capital accumulation comes from (endogenous) gross investment less (exogenous) scrapping. Technical progress is Hicks-neutral. Carbon emissions stem from global output with an emission coefficient which can be reduced by national policies, $\overline{\sigma}_{i,t} = (1 - \mu_{i,t})\sigma_{i,t}$, where $\mu_{i,t} \in (0,1)$ stands for the carbon abatement rate and $\sigma_{i,t}$ is the exogenous carbon intensity of the economy. Abatement costs are given by an increasing and convex cost function $C_i(\mu_{i,t})$. Carbon emissions accumulate in the atmosphere. Concentration, through a simplified carbon cycle, yields a global mean temperature, expressed as temperature change with respect to pre-industrial level, ΔT_t. The impacts of global warming in each country are considered through damage cost functions $D_i(\Delta T_t)$, increasing and convex. Thus, consumption is given by the gross output minus investment, abatement costs and damage costs, $Z_{i,\,t} = Y_{i,\,t} - I_{i,\,t} - C_i(\mu_{i,\,t}) - D_i(\Delta T_t)$. The welfare of each country is measured as the aggregate discounted consumption until the end of the simulation period.

[11] By using the expression of 'Sharing scheme' in the title of their paper, Eyckmans and Finus indicate that they do not propose a particular solution but are interested instead in identifying a class of sharing rules that stabilize all *PIS* coalitions.

[12] Europe is defined as EU-15.

[13] One may find that having 6 regions is too aggregated. This is true for the ROW where identifying some key countries, like India or Brazil would be desirable. But on the other hand it must be noticed that we have the key players, and that more players would make this kind of computational analysis non-manageble. As an example, a 18-region version of the CWS is currently under development: it generates about 270,000 PANEs.

[14] For short, we henceforth use only *country*.

The model is used to determine, over the period 2000–2300, paths of investment (I_{it}) and emissions (through the abatement rate μ_{it}) over time and, consequently, capital accumulation, carbon concentration, temperature change and finally consumption, all at the world and country levels.

This economic model is converted into a six-player dynamic game by letting the six countries be the six players, whose strategies are the decision variables I_{it} and μ_{it}, $\forall i = 1, ..., 6$, $\forall t = 2010, ..., 2300$ (with a 10 year step size), and whose individual payoffs are their respective aggregate discounted consumptions until the end of the period as they result from capital accumulation, carbon concentration and temperature change.

The players-countries' strategies are specified according to a number of alternative scenarios. First, the *Nash equilibrium* scenario,[15] which is the joint outcome of each country maximizing its welfare taking the actions of the others as given. Next, the scenarios called *Partial Agreement Nash Equilibria with respect to a coalition*,[16] each of which is the outcome of a subset of countries maximizing jointly their welfare, while the others act individually (there are as many such scenarios considered as there are coalitions, that is, proper subsets of N). And, finally, the *Pareto efficient* scenario where all countries act jointly so as to maximize the world welfare.

The dynamic optimization problems whose solutions are the numerical values of each one of these scenarios are stated in Appendix.[17] Parameter values as well as initial values are gathered there also. The CWS model allows for different (exogenous) regional discount rates, namely 1.5% in developed countries and 3.0% in developing ones. The huge differences among countries in terms of stage of development and access to financial markets justify this assumption. Higher discount rates for developing countries reflect both a higher degree of impatience and less efficient capital markets.

Finally, transfers between countries are, as in Eyckmans and Tulkens (2003), *generalized GTT transfers*,[18] that is, a dynamic extension due to Germain et al. (1997) of the Chander and Tulkens (1995, 1997) transfers mentioned above.

3.2 Data set and calibration

The CWS model is calibrated on standard international databases. The key data and parameters value are gathered into the Appendix. All details are available in Gerard (2006, 2007). A special attention should be deserved to two key features

[15] In the terminology of dynamic non-cooperative games, this is an 'open loop' Nash equilibrium. 'Closed loop' or 'feedback' Nash equilibria have also been introduced in dynamic core-stability analysis in Germain et al. (2003), albeit with a simpler model. An extension to the CWS model is still awaiting.

[16] These are of open loop nature as well.

[17] The model runs under GAMS. All codes are available from the authors upon request.

[18] The formula, reproduced here as expression A.12 and A.13 in Appendix, is of the same structure as equation (1) in the text above.

that will have a clear influence on model's properties: population profiles and technological changes.

For population growth we use the publications of the United Nations, *World Population to 2300* (2004) and *World Population Prospects: The 2004 Revision* (2005). At this horizon, world population is expected to reach 9 billion people. The time profiles of various regions become are contrasted. Europe, Japan and China face a peak in their population between 2020 and 2030, or even before, and then experience a decline. The population in the Former Soviet Union is expected to decrease while it should be increasing in the USA, mainly because of immigration and fertility rates. In the Rest of the World, short-term population growth would be strong, but followed by a strong slowdown. We assume that, in each country population size converges to a steady state value in the long run.

In the CWS model technological progress encompasses two elements, the global factor productivity and the carbon intensity of economic activity. As far as the former is concerned, high positive trends are expected for China and the USA, while lower progress would occur in Japan, the Former Soviet Union (FSU) and the Rest of the World (ROW). The most striking update concerns carbon intensities which have exhibited contrasting patterns in the recent years. Our data come from the International Energy Agency for carbon emissions and from the World Bank for GDP.[19] Apparently, stringent industrial adjustments are in place that could yield sharp decreases in carbon intensities. This is particularly true for China and FSU. On the contrary, recent trends in Japan and ROW suggest slower carbon improvements.

4. Stability Analysis of Coalitions

We now apply the different concepts of coalition stability to the numerical CWS model. Given the six regions and the 63 coalitions that can possibly form, denoted by S, we compute for each of them its worth W_S in the sense of the gamma-characteristic function, that is, at a Partial Agreement Nash Equilibrium of the model. More precisely, for each S we solve simultaneously the following $n - s + 1$ dynamic optimization problems:

- for the insider, $\forall i \in S : \max W_S = \sum_{i \in S} \sum_{t=0}^{T} \frac{Z_{i,t}}{(1+\rho_i)^t}$
- for the outsider, $\forall i \in N|S :\ \max W_i = \sum_{t=0}^{T} \frac{Z_{i,t}}{(1+\rho_i)^t}$

where each W_i is the value of the objective function A.1 of the CWS model as stated in Appendix, subject to the constraints A.2-A.11.

[19] In fact, we use the *Climate Analysis Indicators Tool* of the *World Resources Institute* that gathers data from the International Energy Agency and the World Bank.

4.1 Core-stability

Let us focus first on the results for the cooperative approach as they appear in Table 1. In this table, the first column contains a six digit key specifying the structure of the coalition: if a region is a member of the coalition, it obtains a '1' at the appropriate position in the key. For instance, the key '111111' refers to $S = N = \{\text{USA, JPN, EU, CHN, FSU, ROW}\}$. Column 2 contains the worth of a coalition (that is the aggregate welfare of its members, $W(S)$) at its corresponding partial agreement Nash equilibrium and column 3 contains the total of what members of each coalition get at the efficient allocation, as achieved by the grand coalition without transfers ($W_S{}^* = \sum_{i \in S} W_i{}^*$). Column 4 gives the difference between the values of the two previous columns. If this difference is negative, it means that S is worse off in the grand coalition. Column 6 gives the total amount of generalized *GTT* transfers for the coalition S ($\Psi_S = \sum_{i \in S} \Psi_i$).

Checking Table 1 reveals two main results. First, without transfers the world efficient allocation, which needs the grand coalition to be achieved, is not core-stable: 18 smaller coalitions (out of 63) can improve upon it. Thus, the grand coalition without transfers cannot form. Second, with *GTT* transfers the world efficient allocation becomes core-stable. This result is of particular importance as it shows that achieving core stability of the world efficient allocation is possible.

Table 1. Coalitions payoffs at all PANE w.r.t. a coalition (W_S^S) and at EFF (W_S^*); generalized GTT transfers (Ψ_S) (billion 1990 US$)

key	$W(S)$	W_S^*	$W_S^*-W(S)$	(%)	Ψ_S	$W_S^*+\Psi_S$	$W_S^*+\Psi_S-W(S)$	(%)
Coalitions of 1 country								
100000	148266	148946	680	0.459	–312	148633	368	0.248
010000	30645	30755	110	0.359	–42	30714	68	0.222
001000	108413	108886	473	0.437	–209	108677	265	0.244
000100	36156	36064	–92	–0.256	196	36260	104	0.288
000010	9745	9790	44	0.454	–23	9766	21	0.217
000001	52326	52107	–219	–0.419	389	52496	170	0.325
Coalitions of 2 countries								
110000	178914	179701	787	0.440	–354	179347	433	0.242
101000	256690	257832	1141	0.445	–521	257311	621	0.242
100100	184488	185009	521	0.283	–116	184893	406	0.220
100010	158016	158735	720	0.455	–335	158400	384	0.243
100001	200852	201052	200	0.100	77	201130	277	0.138
011000	139059	139641	582	0.418	–84	139558	498	0.358
010100	66804	66819	15	0.023	155	66973	170	0.254

(Continued)

Table 1 (Continued)

key	$W(S)$	W_S^*	$W_S^*-W(S)$	(%)	Ψ_S	$W_S^*+\Psi_S$	$W_S^*+\Psi_S-W(S)$	(%)
010010	40391	40544	154	0.381	–65	40480	89	0.220
010001	83016	82862	–154	–0.185	348	83210	194	0.233
001100	144602	144949	348	0.240	–12	144937	335	0.232
001010	118160	118675	515	0.436	–232	118444	283	0.240
001001	160901	160993	92	0.057	181	161173	273	0.170
000110	45902	45853	–49	–0.107	173	46026	124	0.271
000101	88532	88170	–362	–0.409	586	88756	224	0.253
000011	62103	61896	–207	–0.333	366	62263	160	0.257
Coalitions of 3 countries								
111000	287346	288587	1241	0.432	–563	288024	679	0.236
110100	215156	215764	608	0.283	–158	215607	451	0.209
110010	188665	189490	825	0.438	–377	189113	448	0.238
110001	231556	231808	251	0.109	35	231843	287	0.124
101100	293010	293895	885	0.302	–324	293571	560	0.191
101010	266446	267621	1175	0.441	–544	267077	631	0.237
101001	309540	309938	398	0.129	–132	309807	267	0.086
100110	194248	194799	551	0.284	–139	194660	412	0.212
100101	237156	237116	–40	–0.017	274	237389	234	0.098
100011	210630	210842	212	0.101	54	210896	266	0.126
011100	175264	175705	440	0.251	–54	175651	386	0.220
011010	148808	149431	623	0.418	–274	149157	349	0.235
011001	191595	191748	153	0.080	139	191887	292	0.152
010110	76553	76609	56	0.073	132	76740	187	0.245
010101	119214	118926	–289	–0.242	544	119469	255	0.214
010011	92776	92652	–125	–0.134	324	92976	200	0.216
001110	154358	154739	381	0.247	–35	154704	346	0.224
001101	197157	197057	–101	–0.051	377	197433	276	0.140
001011	170672	170782	110	0.065	158	170940	268	0.157
000111	98294	97960	–334	–0.340	563	98522	228	0.232
Coalitions of 4 countries								
111100	323695	324650	956	0.295	–366	324284	590	0.182
111010	297104	298376	1272	0.428	–586	297791	687	0.231
111001	340268	340694	426	0.125	–173	340520	253	0.074
110110	224919	225554	635	0.282	–181	225373	454	0.202
110101	267888	267871	–17	–0.006	232	268103	215	0.080
110011	241338	241597	259	0.107	12	241609	271	0.112
101110	302782	303685	903	0.298	–348	303337	555	0.183
101101	345972	346002	30	0.009	65	346067	95	0.028

(Continued)

Table 1. Coalitions payoffs at all PANE w.r.t. a coalition (W_S^S) and at EFF (W_S^*); generalized GTT transfers (Ψ_S) (billion 1990 US$) (Continued)

key	$W(S)$	W_S^*	$W_S^*-W(S)$	(%)	Ψ_S	$W_S^*+\Psi_S$	$W_S^*+\Psi_S-W(S)$	(%)
101011	319333	319728	395	0.124	–155	319573	240	0.075
100111	246948	246905	–43	–0.017	250	247156	208	0.084
011110	185022	185494	472	0.255	–77	185417	395	0.213
011101	227875	227812	–64	–0.028	335	228147	272	0.119
011011	201370	201538	168	0.083	116	201653	283	0.141
010111	128982	128715	–267	–0.207	521	129236	254	0.197
001111	206940	206846	–94	–0.046	354	207200	260	0.125
Coalitions of 5 countries								
111110	333468	334440	971	0.291	–389	334051	582	0.175
111101	376733	376757	24	0.006	23	376780	47	0.012
111011	350063	350483	420	0.120	–196	350287	223	0.064
110111	277685	277661	–25	–0.009	209	277869	184	0.066
101111	355782	355791	9	0.003	42	355833	51	0.014
011111	237663	237601	–62	–0.026	312	237913	251	0.105
Coalitions of 6 countries								
111111	386547	386547	0	0.000	0	386547	0	0.000

4.2 Internal-external stability

Table 2 presents the results for the non-cooperative approach. The columns refer, for the various coalitions, to the three different stability properties [internal (*IS*), external (*ES*), and potential internal (*PIS*)] proposed by this approach. A cross in a column means that the property is satisfied for the corresponding coalition. We summarize the main results as follows, distinguishing again between without and with transfers cases:

- Internal and external stability: very few coalitions pass the *IS* test (8 or 7 of them, out of 57).[20] In particular, the grand coalition, that is, the one that would achieve the world efficient allocation without transfers, does not pass it. More coalitions (11 or 15 out of 56 – the grand coalition is irrelevant here) pass the *ES* test. No coalition passes both tests however, except for one, namely the couple USA, EU.
- Potential internal stability: contrary to the *IS* and *ES* tests, the PIS test is one that implicitly refers to transfers within the coalitions, with the purpose of inducing internal stability. Here again, the grand coalition does not pass the

[20] Here we exclude singletons.

Table 2. Non cooperative stability properties satisfied by different coalitions

Coalition	*IS*	*ES*	*PIS*
USA, JPN			✓
USA, EU	✓	✓	✓
USA, CHN			✓
USA, FSU			✓
USA, ROW			✓
JPN, EU			✓
JPN, CHN			✓
JPN, FSU			✓
JPN, ROW	✓		✓
EU, CHN			✓
EU, FSU			✓
EU, ROW			✓
CHN, FSU	✓		✓
CHN, ROW	✓		✓
FSU, ROW	✓		✓
USA, JPN, EU		✓	✓
USA, JPN, CHN			✓
USA, JPN, FSU			✓
USA, JPN, ROW			✓
USA, EU, CHN		✓	✓
USA, EU, FSU		✓	✓
USA, EU, ROW		✓	✓
USA, CHN, FSU			✓
USA, CHN, ROW			✓
USA, FSU, ROW			✓
JPN, EU, CHN			✓
JPN, EU, FSU			✓
JPN, EU, ROW			✓
JPN, CHN, FSU			✓
JPN, CHN, ROW			✓
JPN, FSU, ROW	✓		✓
EU, CHN, FSU			✓
EU, CHN, ROW			✓
EU, FSU, ROW			✓
CHN, FSU, ROW	✓		✓
USA, JPN, EU, CHN		✓	
USA, JPN, EU, FSU		✓	✓
USA, JPN, EU, ROW		✓	
USA, JPN, CHN, FSU			✓
USA, JPN, CHN, ROW			✓
USA, JPN, FSU, ROW			✓
USA, EU, CHN, FSU		✓	
USA, EU, CHN, ROW		✓	
USA, EU, FSU, ROW		✓	
USA, CHN, FSU, ROW			✓
JPN, EU, CHN, FSU			✓
JPN, EU, CHN, ROW			✓
JPN, EU, FSU, ROW			✓
JPN, CHN, FSU, ROW			✓
EU, CHN, FSU, ROW			✓
USA, JPN, EU, CHN, FSU		✓	
USA, JPN, EU, CHN, ROW		✓	
USA, JPN, EU, FSU, ROW		✓	
USA, JPN, CHN, FSU, ROW			✓
USA, EU, CHN, FSU, ROW		✓	
JPN, EU, CHN, FSU, ROW			
GRAND COALITION		irrelevant	

Notes: *IS* = Internal Stability,
ES = External Stability,
PIS = Potential Internal Stability.
✓ means that the property is satisfied for the coalition.

test, and only 1 five-country coalition passes the test. However, many smaller coalitions do. More precisely, 10 four-country coalitions, out of 15, are *PIS*, and all the three-country and two-country coalitions are. In sum, only 5 coalitions (out of 63) are not *PIS*.

These results are in line with the main conclusion of the theoretical literature on *IS-ES* stability,[21] namely that no large coalitions can be stable in that sense. There is however the following novel interest with the present results: as this theoretical literature establishes its claim only for simple models with identical countries, it is shown here by an example that the thesis may also holds by and large in the case of a much more complex economic model and for non identical countries. On the question whether transfers can improve that stability, our mostly negative results do also confirm those obtained by Eyckmans and Finus (2004) and Carraro et al. (2006).

4.3 Core and internal-external stability compared

Considering the grand coalition N, we can report the following three results:

1. Without transfers, the world efficient allocation, that only the grand coalition can achieve, is lacking stability in both the core sense and the internal-external sense when computed with the *CWS* model.
2. By contrast, if transfers are introduced, the world efficient allocation achieved by N can be stabilized in the core sense, by means of *GTT* transfers within the grand coalition.
3. This is not possible in the internal-external sense, i.e. by means of *PIS* transfers.

The reason for this difference (*GTT* transfers work while *PIS* transfers do not) is in the logic that lies behind the two stability concepts: in the core case, stability of N is obtained from threatening the objecting parties to be deprived of any part in the surplus generated by the collective move to efficiency. By construction, this is always feasible. In the internal-external stability case, stability results from offering each country its free rider payoff; but there is no general assurance that this be always feasible: the surplus generated by the move to efficiency may be insufficient for ensuring that payoff to *all* countries. This depends upon characteristics of the computational model, such as, e.g. the distance in welfare terms between the Nash and Pareto solutions, that is, the size of the surplus.

As far as coalitions other than N are concerned, none of them can evidently be stable in the core sense because it is precisely the meaning of the core result that N with transfers can improve upon any of them. Concerning their stability in the internal-external stability sense, one finds in Tables 1 and 2 hardly any correlation between those coalitions that meet either internal or external stability (coalitions with an '✓' in the *IS* or *ES* columns of Table 2) and those which could block in the core sense the efficient allocation without transfers (coalitions with a negative sign

[21] As initiated by Barrett (1994) and Carraro and Siniscalco (1993); Asheim et al. (2006) is in the same spirit.

in column 4 of Table 1). In short, this is because *the reasons for blocking* (which are, for the members of S, the hope to do better by themselves) *are fundamentally different from those for free riding* (which are the search for benefit from the others' actions). This last argument also explains why the *PIS* property prevails better with small coalitions: *vis-à-vis* a small coalition, there is little to free ride about (because the coalition does not achieve much), so that the surplus generated can be sufficient to deter from such behavior.

In summary, the core *vs* internal-external stability concepts have quite opposing properties, not only as to the grand coalition, N, but also for smaller ones. One concept excludes small coalitions, whereas the other concept can be found to be satisfied with small coalitions.

5. Stability *versus* Performance

Can policy implications be derived from the above stability discussion and simulation results? In particular, how important are the coalitional stability properties we have identified? Should they serve as an argument to support or advocate specific structures for climatic international agreements such as small coalitions rather than large ones, or homogeneous rather than heterogeneous ones?

To answer these questions, let us consider two criteria measuring the global outcome resulting from an agreement, that is,

- the aggregate welfare level reached at the world level,
- the environmental performance achieved, expressed by atmospheric carbon concentration.

and consider how these are met by alternative coalition structures. This is done in Figure 1 with the numerical results provided by the CWS model. On the two axes we use a welfare and an environmental index respectively, that we borrow from CEF-06. Both indexes give the value 1 to the world efficient allocation (the grand coalition case) that produces the highest aggregate welfare and the lowest carbon concentrations, and the value 0 to the non-cooperative Nash case, that depicts the lowest aggregate welfare and the highest carbon concentrations. Formally, the indexes are computed as follows:

Welfare index: $$I^W(S) = \frac{\sum_{i\in N}(W_i(S) - W_i^{Nash})}{\sum_{i\in N}(W_i^* - W_i^{Nash})},$$

Environmental index: $$I^E(S) = \frac{M_{2300}^{Nash} - M_{2300}(S)}{M_{2300}^{Nash} - M^*{}_{2300}},$$

where $\sum_{i\in N} W_i(S)$ and $M_{2300}(S)$ are respectively the aggregate welfare and carbon concentration levels in 2300 under the corresponding coalition structure *S*, while '*' refers to the world efficient allocation (full cooperation) and 'Nash' refers to the Nash case (no cooperation). An increasing relation is obtained with the non-cooperative Nash equilibrium (lowest global welfare, highest carbon concentration) at the bottom left and the grand coalition (highest global welfare, lowest carbon concentration) at the top right.

Remembering that internal stability in its potential form prevails with small coalitions while core-stability is achieved only with the largest one, the relation also depicts both the welfare and the environmental performances of alternative coalition sizes.

Figure 1 displays many appealing results. First, it shows that different coalitions are able to provide similar outcome, either for welfare or environmental quality. Put differently, an improvement in the environmental quality does not necessarily goes with an improvement in welfare at the world level, and conversely. The outcome depends on the coalition. As an example, it is striking to see that a coalition formed by three countries, namely {CHN, FSU, ROW}, performs as well as a 5-country coalition in terms of environmental quality, namely {USA, JPN, EU, CHN, FSU}. Still, the former ranks much higher in terms of global welfare. It shows that a smaller coalition may perform better than a larger coalition. This result is even reinforced by the fact that the former coalition is internally stable while the latter cannot be stabilized. Another striking result is the performance of the Annex B coalition: it is almost similar to the Nash equilibrium.

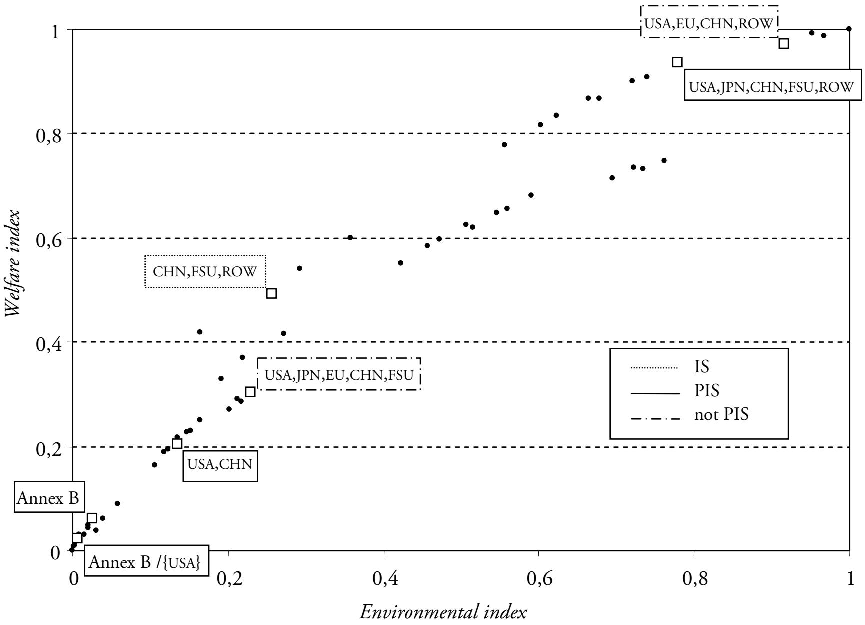


Figure 1. Global outcome (aggregate welfare and the environment) with alternative coalition structures

Finally, two coalitions are of special interest because of their performance: {USA, EU, CHN, ROW} and {USA, JPN, CHN, FSU, ROW}. The former is quite close to the grand coalition. The latter is almost at the same welfare performance level but with a somewhat lower environmental index. How can this be explained? First, it must be noticed that the former can not be stabilized, while the latter is PIS. In other words, the former cannot form, while the latter can because it is beneficial to all parties. This makes a huge difference between the two in terms of political applicability. Second, the latter enlarges the coalition by inviting FSU and JPN, but puts the EU outside. By doing so, it makes the coalition PIS. In our model (as well as in many integrated models) the EU is known to have large climate damages. As a consequence, it asks for strong carbon emission reductions, which is costly for all coalition members. By putting the EU outside and inviting FSU and JPN, the coalition becomes potentially internally stable, the world welfare level is almost the same and the climate is better-off.

Clearly, accepting or recommending small coalition arrangements because of their potential internal stability virtues entails a loss on both counts, that striving for an efficient and core stable alternative could avoid. Internal stability thus appears to be a weakly desirable objective.

6. Is Coalition Homogeneity Desirable?

A common argument in the climate policy debate is that developed countries should engage themselves first, and developing countries would thereafter be invited to join the agreement and participate in the mitigation process. Although this argument seems reasonable on the ground of historical responsibilities, one may question its effectiveness in combating climate change.[22] This question has been partly addressed by McGinty (2007) who shows that the benefits from cooperation are greater when countries are heterogenous. Here, we go one step further by linking effectiveness with stability. We shall analyze how the composition of a coalition, that is, its degree of *homogeneity* (which is to be defined), affects its stability.

The regions/countries considered in the CWS model can be split into two categories:

- developed-Annex B countries (USA, EU and JPN), with high per capita emissions and GDP,
- developing-non-Annex B countries (CHN and ROW), with low per capita emissions and GDP, and low-cost abatement opportunities.

[22] This is the principle of 'common but differentiated responsibilities' of countries stated in the UN Framework Convention.

In the following we will talk about an *heterogeneous coalition* when a coalition is formed by countries coming from more than a single category. Conversely, an *homogeneous coalition* will designate a coalition formed by countries from a single category. The FSU will move as a free electron in this categorization as it offers the characteristics of both a developed country (high emissions per capita) and a developing one (low cost abatement opportunities, low GDP per capita). Accordingly, our 57 coalitions (excluding singletons) are broken down into 42 heterogeneous coalitions and 15 homogeneous ones. We examine the relation mentioned above, successively without and with transfers

In the no transfer case, all the 4 homogeneous coalitions involving FSU and developing-non-Annex B countries pass the *IS* test, and the homogenous coalition {USA, EU} is both internally and externally stable. On the other hand, 5 of the 7 internally stable coalitions are homogenous coalitions. Among these 5 homogenous IS coalitions, only one involve developed countries, USA or EU. The two heterogenous IS coalitions include JPN as developed-Annex B country, which is the least important emitter of the six regions.[23] So it seems that adding a large developed country to an homogenous coalition of developing country is detrimental to its internal stability.

It is sometimes argued that, for the sake of effectiveness, the big polluters of each category should be included in a coalition. In CWS, the two main polluters in each category are USA or EU, on the one hand, and CHN or ROW on the other hand. It appears that none of the coalitions involving at least one of these big polluters is internally stable. Moreover, none of the coalitions that involve the two main emitters of a category and at least one emitter of the other category is internally stable.

When the possibility of transfers is introduced, again stability seems to be enhanced by homogeneity. Indeed, it is striking to see that the 5 coalitions that are not PIS are all heterogenous ones. Those coalitions are large, as they gather 4 or 5 countries. Put differently, all the homogenous coalitions can be stabilized, but those coalitions are smaller. Interestingly, the Annex B coalition turns out to be more stable than the 'Annex B without the USA' coalition.[24] Indeed, this latter coalition does not satisfy the *ES* property: this means that the United States would be better off by coming back to the Annex B coalition. Furthermore, no four-country (or more) coalitions that involve both the USA and the EU and at least one non-Annex B countries pass the *PIS* test.

The discussion about homogeneity vs heterogeneity can also be analyzed by using Figure 1. One can see that the 'best' (in terms of global welfare) homogeneous coalition, namely {CHN, FSU, ROW}, leads to far lower global welfare

[23] JPN is less important in terms of emissions than USA or EU.

[24] The so-called *Present Kyoto* coalition in CEF-06.

and far higher carbon concentrations than both the 'best' heterogeneous coalition (the grand coalition) and the 'best' heterogeneous coalition satisfying the PIS property, that is, {USA, JPN, CHN, FSU, ROW}. As a consequence, promoting homogeneous coalitions would lead to very low mitigation policies at the world level, unable to tackle climate change issue as heterogeneous (larger) coalitions could do.

In sum, there seems to be a trade-off between stability and environmental effectiveness. Homogeneity in climate coalitions fosters stability but is detrimental to climate effectiveness.

7. Sensitivity Analyses

The objective of this section is to test to what extent our results are robust to the choice of some key parameters. Extensive sensitivity analyses have revealed that two assumptions may be key (Gerard, 2006). The first one is the evolution of carbon intensity (σ_{it} in equations of Appendix) in China in the forthcoming years, and the second one is the slope of the damage functions in all countries. They will be considered in the two first sub-sections. Then, we will pay some attention to the update of the CWS model, in particular in terms of carbon intensity profiles and population profiles between the version used in ET-03 and the current one. The question here is to see if updating the economic part of such a the model can alter our conclusions or not. This will be done in a last subsection. Sensitivity analyses with respect to the discount rate have not revealed important varying results as to the stability of alternative coalitions with respect to this parameter.

7.1 Carbon intensity in China

China is expected to become the world largest carbon emitter soon, but when heavily depends on the assumption made on technological progress. In our model, carbon intensity and total factor productivity are calibrated and projected on the basis of past profiles, which yields a quite rapid – and optimistic – decarbonization of the Chinese economy in the forthcoming decades. As a first sensitivity analysis, we reduced the rate of decarbonization by half, while keeping the asymptotical value unchanged. This raises Chinese emissions by 60% in the *business-as-usual* scenario in 2100 while the level of emissions in the very long-term is kept unchanged. The fact that Chinese emissions are higher increases the climate externality generated (the effect of its own strategy on the other countries) and therefore the possible gain from cooperation. However, the free-riding incentive may also be stronger for the other countries in the coalitions including China because these coalitions will internalize a larger part of the global externality. Both effects potentially raise concern for stability.

The model shows that the gain in world welfare between the Nash equilibrium and the efficient scenarios is slightly increased by around 1%. Our main results on the core-stability of the grand coalition and the best *PIS* coalition (which includes China) still prevail. The effect on the stability of coalitions without China is negative: the difference between the aggregate welfare of the coalition and the sum of the free-riding claims of its members (definition of the *PIS* property) decreases for 23 out of the 26 coalitions considered; indeed, such coalitions internalize a smaller part of the externality. However, the effect on the coalitions including China is less clear: it increases for 16 out of 31 coalitions, but decreases for 18. In short, the model confirms the mechanisms at stake in this test and our main conclusions remain valid. The surprise may be that the effect on global welfare gain from cooperation is quite low.

7.2 Slope of damage functions

The second sensitivity analysis concerns the damage functions. These, still borrowed from Nordhaus and Yang (1996), bear major uncertainties. The relationship between global temperature increase and climatic impacts is highly difficult to quantify, and the most recent studies (including the Stern Review and the Fourth IPCC Assessment Report) seem to suggest higher damage sensitivity. We did this by increasing the exponent of the damage functions ($\theta_{i,2}$ in equations of Appendix) by 50% in all countries. Intuitively, this will reinforce the climate externality, and thus the desirability of cooperation. But, it is difficult to infer, *a priori*, the implication for stability because the free-riding incentive may also be stronger when the coalitions try to better internalize the climate externality.

After computation the CWS model confirms that the gain in global welfare associated with cooperation is stronger, and this time the increase is significant (the gain is three times higher). However, even with such a strong incentive for cooperation, our main results on core-stability of the grand coalition and the best *PIS* coalition remain valid. This means that the stronger gain from cooperation dominates the reinforcement of the free-riding incentives. No clear conclusion can be drawn about the impact on the stability of the other coalitions. Indeed, the difference between the aggregate welfare of the coalition and the sum of the free-riding claims of its members increases for 38 out of 57 coalitions, but decreases for 19 others, making 6 coalitions no more *PIS*. The increase concerns mainly small coalitions, for which we have already mentioned that there is less to free-ride about.

7.3 Economic update

In this paper we use an updated version of the CWS model initially presented in ET-03.[25] The update consists essentially in changes in the numerical value of several

[25] The details of this update are reported in the discussion paper version of our article, Bréchet et al. (2007).

parameters of the optimization model (A.1)-(A.11), reflecting new assumptions on population growth and technological change. These have two main implications for the scenarios. First, world emissions are lower in the business-as-usual scenario than they were in the previous version of the model. Second, heterogeneity among countries is reinforced: national emission profiles are generally lower in all countries, in particular in China, but the USA experience higher emissions. Thus, the relative weight of countries in the global system is significantly changed, and so do the costs and benefits for each country of participating in a given climate agreement. The implications for our coalitional stability analyzes are as follows.

About the cooperative approach, the main economic theoretic point is to verify whether a gamma-core solution can also be found with the new values of the parameters, as was the case with the original ones.[26] The result happens to be positive. Here, as in the previous version, GTT transfers need to be used because, without them, the efficient solution is blocked by 18 coalitions (a number that was 14 previously). The concept of gamma-core thus appears to be robust to our updating. But the presence of four newly blocking coalitions may be seen as revealing an increased instability of the efficient allocation without transfers. This makes the transfers all the more necessary if efficiency is being sought in the international agreement.

As far as the non-cooperative approach is concerned, in both versions of CWS very few coalitions are internally stable (8 or 7 of them, out of 57). A few more coalitions (11, or 15, out of 56) are externally stable. No coalition passes both tests, except the couple {USA, EU} which does so only in the updated version. When transfers are introduced, 2 three-country coalitions that were not stable in the first version become potentially internally stable (PIS) after the update, namely {USA, EU, CHN} and {JPN, CHN, FSU}. The number of four-country coalitions that are PIS remains the same in the two versions (10, out of 15).

Finally, as to the distinction between homogenous vs. heterogenous coalitions in relation with stability, we find that without transfers, while 6 of the 8 internally stable coalitions were heterogeneous coalitions in the earlier version, only two of these 6 heterogeneous coalitions still pass the *IS* test after the update. With transfers, homogeneity favors somewhat more the stability of coalitions in the updated version of CWS than in the original one.

8. Conclusion and Policy Implications

In the literature on international climate agreements, two alternative game theoretic approaches are used to discuss the stability of climate coalitions, which are based on two different stability concepts, namely ‘gamma-core’ stability and

[26] Remember that existence of a gamma-core solution is established analytically only for the usual basic models (linear and convex, respectively) of Chander and Tulkens 1995, 1997, not for the CWS model.

‘internal-external’ stability. With the integrated assessment CWS model, this paper numerically compares and contrasts the results obtained from applying to it either one of these approaches. From a methodological viewpoint, it turns out that, in this model, transfers are required to ensure the stability of most coalitions whatever the concept used. But transfers are not equally successful to stabilize coalitions in either approaches because of the different logic that lies behind the two concepts. More precisely, while transfers can make the grand coalition stable in the gamma-core sense (which rests on the threat of failing to reach an agreement), this is never the case in the internal-external stability sense (which rests on offering compensation for resisting the temptation of free riding); only smaller coalitions, where there is little to free-ride about, are found stable in this sense, sometimes with transfers. Moreover, while we note that homogeneity among the members of a coalition appears to help the coalition’s potential internal stability irrespective of its size, the global outcome in terms of either aggregate welfare or environmental performance reached by small or homogeneous coalitions is far less attractive compared with the world efficient allocation that what can be reached by the heterogeneous grand coalition only.

Policy-wise, these results bring strong support to the view that environmental agreements which include a large number of countries are desirable both in terms of the countries’ welfare as in terms of global environmental performance. In addition, stability in the gamma-core sense can be achieved only if the agreement includes all countries of the world, whereas stability in the internal-external sense can be achieved only among smaller numbers of signatories. Therefore, agreements including all countries, such as the Kyoto Protocol (before the withdrawal of the USA), are most desirable from the three points of view of welfare, environment, and stability.

As illustrated in the paper, the last property can be ensured by means of appropriately designed transfers of resources. These can take many forms, some of which are quite different from the lump sum ones used here. Among them, and most importantly, the transfers implied by a *cap and trade* scheme of the type established by the Kyoto Protocol do have all the stability properties required here for transfers – and a few more virtues as well.[27]

Finally, if for reasons other than those invoked above, a treaty involving the ‘grand’ coalition of all countries cannot be signed and smaller coalitions are envisaged, the above simulations indicate that heterogeneity of composition matters more than size for the stability of a coalition.[28] Thus, promoting homogeneous coalitions, as is sometimes done, is not supported by our analysis if effectiveness is taken as a policy objective.

[27] For a full development of this point, which is often overlooked, see Chander et al. (2002). For an analysis applied to the EU unilateral strategy before Copenhagen, see Bréchet et al. (2010).

[28] See Bréchet and Eyckmans (2010) for further analyzes about this point.

Appendix

Statement of the CWS model. The index $i = 1,\ldots n$ stands for region/country.

Objective functions

$$W_i = \sum_{t=0}^{T} \frac{Z_{i,t}}{(1+\rho_i)^t} \tag*{A.1}$$

Constraints

$$Y_{i,t} = A_{i,t} K_{i,t}^{\alpha} L_{i,t}^{1-\alpha} \tag*{A.2}$$

$$Y_{i,\,t} = Z_{i,\,t} + I_{i,\,t} + C_i(\mu_{i,\,t}) + D_i(\Delta \mathrm{T}_t) \tag*{A.3}$$

$$K_{i,\,t+1} = (1-\delta_K)^{10} K_{i,\,t} + \ 10 I_{i,\,t}, \text{ with } K_{i,0} \text{ given} \tag*{A.4}$$

$$\mathrm{E}_{i,\,t} = \sigma_{i,\,t}(1-\mu_{i,\,t}) Y_{i,\,t} \tag*{A.5}$$

$$C_i(\mu_{i,\,t}) = Y_{i,t} b_{i,1} \mu_{i,t}^{b_{i,2}} \tag*{A.6}$$

$$M_{t+1} = \overline{M} + \beta \sum_{j=1}^{n} E_{j,t} + (1-\delta_M)(M_t - \overline{M}), \text{ with } M_0 \text{ given} \tag*{A.7}$$

$$F_t = 4.1\ ln(M_t/M_0)/ln(2) \tag*{A.8}$$

$$T_t^0 = T_{t-1}^0 + \tau_3\ (\Delta T_{t-1} - T_{t-1}^0), \text{ with } T_0^0 \text{ given} \tag*{A.9}$$

$$\Delta T_t = \Delta T_{t-1} + \tau_1(F_t - \lambda \Delta T_{t-1}) - \tau_2(\Delta \mathrm{T}_{t-1} - T_{t-1}^0), \text{ with } \Delta T_0 \text{ given} \tag*{A.10}$$

$$D_i(\Delta \mathrm{T}_t) = Y_{i,\,t}\theta_{i,1}(\Delta T_t/2.5)^{\ \theta_{\mathrm{i},2}} \tag*{A.11}$$

Solutions

Pareto efficient: $(\mu^*_{i,\,t}, I^*_{i,\,t})_{\substack{i=1,\ldots,n \\ t=0,\ldots,T}}$ that solves:

$$\mathrm{Max} \sum_{t=0}^{T} \sum_{i=1}^{N} (\mathrm{A.1}) = \sum_i W_i^*, \text{ subject to (A.2)...(A.11).}$$

Nash equilibrium: $(\mu_{i,t}^{NE}, I_{i,t}^{NE})_{\substack{i=1,\ldots,n \\ t=0,\ldots T}}$ that solves, for each $i = 1,\ldots, n$:

$$\mathrm{Max} \sum_{t=0}^{T} (\mathrm{A.1}) = \sum_i W_i^{NE} \text{ subject to (A.2)...(A.11), with } E_{j,t} = E_{j,t}^{NE}, \forall j \neq i,\ t = 0,\ldots\ T.$$

Partial Agreement Nash equilibria w.r.t. any coalition $S \in N$:

$(\mu_{i,t}^{S}, I_{i,t}^{S})_{\substack{i=1,\ldots,n \\ t=0,\ldots T}}$ that solves:

$$\mathrm{Max} \sum_{t=0}^{T} \sum_{i=1}^{n} (\mathrm{A.1}) = \sum_i W_i^{S} \text{ subject to (A.2)...(A.11) with } E_{j,t} = E_{j,t}^{S}, \forall j \notin S,$$

$t = 0,\ldots$ T, and $\forall i \notin$ S,

$$\mathrm{Max} \sum_{t=0}^{T} (\mathrm{A.1}), \text{ subject to (A.2)...(A.11) with } E_{j,t} = E_{j,t}^{S}, \forall j \neq i,\ t = 0,\ \ldots,\ T.$$

GTT transfers

$$\Psi_i = -(W_i^* - W_i^{NE} + \pi_i(\textstyle\sum_{j\in N} W_j^* - \sum_{j\in N} W_j^{NE}) \qquad \text{A.12}$$

$$\pi_i = \frac{\sum_{t=0}^{T} D_t'(\Delta T_i^*)/(1+\rho_i)^t}{\sum_{j\in N}\sum_{t=0}^{T} D_j'(\Delta T_i^*)/(1+\rho_j)^t} \qquad \text{A.13}$$

Table I: List of variables

$Y_{i,t}$	Production (billions 1990 US$)
$A_{i,t}$	Productivity
$Z_{i,t}$	Consumption (billions 1990 US$)
$I_{i,t}$	Investment (billions 1990 US$)
$K_{i,t}$	Capital stock (billions 1990 US$)
$L_{i,t}$	Population (million people)
$C_{i,t}$	Cost of abatement (billions 1990 US$)
$D_{i,t}$	Damage from climate change (billions 1990 US$)
$E_{i,t}$	Carbon emissions (billions tons of C)
$\sigma_{i,t}$	Carbon intensity of GDP (kgC/1990 US$)
$\mu_{i,t}$	Carbon emission abatement rate
M_t	Atmospheric carbon concentration (billions tons of C)
F_t	Radiative forcing (Watt per m²)
ΔT_t	Temperature increase atmosphere (°C)
T_t^o	Temperature increase deep ocean (°C)
W_i	Welfare (billions 1990 US$)

Table II: Global parameter values

δ_K	Capital depreciation rate	0.10
γ	Capital productivity parameter	0.25
β	Airborne fraction of carbon emissions	0.64
δ_M	Atmospheric carbon removal rate	0.08333
τ_1	Parameter temperature relationship	0.226
τ_2	Parameter temperature relationship	0.44
τ_3	Parameter temperature relationship	0.02
λ	Feedback parameter	1.41
$\overline{M}$	Pre-industrial carbon concentration	590
M_0	Initial carbon concentration in 2000	783
ΔT_0	Initial temperature change atmosphere in 2000	0.622
T_0^0	Initial temperature change deep ocean in 2000	0.108

Table III: Regional parameter values

	$\theta_{i,1}$	$\theta_{i,2}$	$b_{i,1}$	$b_{i,2}$	ρ_i
	Damage function		Abatement cost function		Discount rate
USA	0.01102	2.0	0.07	2.887	0.015
JPN	0.01174	2.0	0.05	2.887	0.015
EU	0.01174	2.0	0.05	2.887	0.015
CHN	0.01523	2.0	0.15	2.887	0.030
FSU	0.00857	2.0	0.15	2.887	0.015
ROW	0.02093	2.0	0.10	2.887	0.030

Table IV: 2000 reference year variables

	$Y_{i,0}$	(%)	$K_{i,0}$	(%)	$L_{i,0}$	(%)	$E_{i,0}$	(%)
USA	7563.8099	27.45	19740.6885	27.97	282.224	4.66	1.5738	24.01
JPN	3387.9305	12.29	9753.9695	13.82	126.870	2.10	0.3295	5.03
EU	8446.9010	30.65	22804.4771	32.31	377.136	6.23	0.8875	13.54
CHN	968.9064	3.52	2686.0563	3.81	1262.645	20.86	0.9468	14.44
FSU	558.4360	2.03	1490.0376	2.11	287.893	4.76	0.6258	9.55
ROW	6633.4274	24.07	14105.2089	19.98	3715.663	61.39	2.1918	33.44
World	27559.4112	100.00	70580.4379	100.00	6052.4310	100.00	6.5552	100.0
	billion 1990 US$	(%)	billion 1990 US$	(%)	million people	(%)	billion tons of carbon (GtC)	(%)

References

Asheim, G., C.B. Froyn, J. Hovi and F.C. Menz (2006), ‘Regional versus global cooperation on climate control’, *Journal of Environmental Economics and Management* **51**(1), 93–109.

Barrett, S. (1994), ‘Self-enforcing international environmental agreements’, *Oxford Economic Papers* **46**, 804–878.

Barrett, S. (2003), *Environment and Statecraft: The Strategy of Environmental Treaty-Making*, Oxford: Oxford University Press.

Bernard, A., A. Haurie, M. Vielle and L. Viguier (2008), ‘A two-level dynamic game of carbon emission trading between Russia, China, and Annex B countries’, *Journal of Economic Dynamics and Control* **32**(6), 1830–1856.

Bréchet, Th., F. Gerard and H. Tulkens (2007), ‘Climate coalitions: A theoretical and computational appraisal’, CORE discussion paper 2007/3 and FEEM Nota Di Lavoro 21.2007.

Bréchet, Th. and J. Eyckmans (2010), ‘Coalition theory and integrated assessment modeling: Lessons for climate governance’, in E. Brousseau, P.A. Jouvet and T. Tom Dedeurwareder (eds), *Governing Global Environmental Commons: Institutions, Markets, Social Preferences and Political Games*, Oxford: Oxford University Press.

Bréchet Th., J. Eyckmans, F. Gerard, Ph. Marbaix, H. Tulkens and J.-P. van Ypersele (2010), 'The impact of the unilateral EU commitment on the stability of international climate agreements', *Climate Policy* **10**, 148–166.

Carraro, C. and D. Siniscalco (1993), 'Strategies for the international protection of the environment', *Journal of Public Economics* **52**(3), 309–328.

Carraro, C., J. Eyckmans and M. Finus (2006), 'Optimal transfers and participation decisions in international environmental agreements', *Review of International Organisations* **1**(4), 379–396.

Chander, P. and H. Tulkens (1995), 'A core-theoretic solution for the design of cooperative agreements on transfrontier pollution', *International Tax and Public Finance* **2**(2), 279–294. Reprinted as chapter 5 in A. Ulph (ed.) (2001), *Environmental Policy, International Agreements, and International Trade*, Oxford: Oxford University Press, pp. 81–96.

Chander, P. and H. Tulkens (1997), 'The core of an economy with multilateral environmental externalities', *International Journal of Game Theory* **26**(3), 379–401.

Chander, P., H. Tulkens, J.-P. van Yperseleand and S. Willems (2002), 'The Kyoto Protocol: An economic and game theoretic interpretation', in B. Kriström, P. Dasgupta and K.-G. Löfgren (eds), *Economic Theory for the Environment: Essays in Honor of Karl-Göran Mäler*, chapter 6, Cheltenham: Edward Elgar, pp. 98–117.

Chander, P., J. Drèze, C.K. Lovell and J. Mintz (eds) (2006), *Public Goods, Environmental Externalities and Fiscal Competition: Essays by Henry Tulkens*, New York: Springer.

Chander, P. and H. Tulkens (2006), 'Cooperation, stability and self-enforcement in international environmental agreements: A conceptual discussion' in R. Guesnerie and H. Tulkens (eds.), *The Design of Climate Policy*, CESifo Seminar Series, Boston: The MIT Press.

Chander, P. (2007). 'The gamma-core and coalition formation', *International Journal of Game Theory* **35**(4), 539–556.

D'Aspremont, C., A. Jacquemin, J.J. Gabszewicz and J.A. Weymark (1983), 'On the stability of collusive price leadership', *Canadian Journal of Economics* **16**(1), 17–25.

D'Aspremont, C., A. Jacquemin and J.J. Gabszewicz (1986), 'On the stability of collusion', in J.E. Stiglitz and G.F. Mathewson (eds), *New developments in the analysis of market structure*, chapter 8, Cambridge, MA: The MIT Press, pp. 243–264.

Eyckmans, J. and M. Finus (2004), 'An almost ideal sharing scheme for coalition games with externalities', CLIMNEG Working Paper 62, Leuven, Belgium: Katholieke Universiteit Leuven.

Eyckmans, J. and H. Tulkens (2003), 'Simulating coalitionally stable burden sharing agreements for the climate change problem', *Resource and Energy Economics* **25**, 299–327.

Gerard, F. (2006), 'Formation des coalitions pour les négociations climatiques internationales: Une approche par le modèle CWS', Master's degree thesis, Department of Economics, Université catholique de Louvain.

Gerard, F. (2007), 'CWS 1.2, une mise à jour du modèle CWS', CLIMNEG Working Paper 80, CORE, Université catholique de Louvain.

Germain, M., P.L. Toint and H. Tulkens (1997), 'Financial transfers to ensure international optimality in stock pollutant abatement', in F. Duchin, S. Faucheux, J. Gaudy and I. Nicolaï (eds.), *Sustainability and firms: technological change and changing regulatory environment*, Celtenham, UK and Brookfield, US: Edward Elgar.

Germain, M., P.L. Toint, H. Tulkens and A. de Zeeuw (2003), 'Transfers to sustain dynamic core-theoretic cooperation in international stock pollutant control', *Journal of Economic Dynamics and Control* **28**, 79–99.

Nordhaus, W.D. and Z. Yang (1996), 'A regional dynamic general-equilibrium model of alternative climate-change strategies', *American Economic Review* **86**(4), 741–765.

McGinty, M. (2007), 'International environmental agreements among asymmetric nations', *Oxford Economic Papers* **59**(1), 45–62.

Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat (2004), *World Population to 2300*, New York: United Nations Publication.

Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat (2005), *World Population Prospects: The 2004 Revision*, New York: United Nations Publication.

Tulkens, H. (1998), 'Cooperation vs. free riding in international environmental affairs: Two approaches', in N. Hanley and H. Folmer (eds), *Game Theory and the Environment*, Chapter 2, 30–44, Cheltenham: Edward Elgar.

World Resources Institute (2007), 'Climate analysis indicators tool (CAIT) version 4.0', Available at http://cait.wri.org.

Yang, Z. (2008), *Strategic Bargaining and Cooperation in Greenhouse Gas Mitigations – An Integrated Assessment Modeling Approach*, Cambridge, MA and London, UK: MIT Press.

VOLUMES PUBLISHED BY FONDAZIONE ENI ENRICO MATTEI
IN THE FEEM PRESS SERIES

The Global Revolution of Unconventional Oil: New Markets, New Governances, New Policies, by John M. Deutch, Milan, FEEM Press, Economy and Society Series, 1/2014

Islamism and Modernity: an Unconventional Perspective, by Tarek Heggy, Milan, FEEM Press, Economy and Society Series, 2/2014

The Ukraine Debacle, by Anatol Lieven, Milan, FEEM Press, Economy and Society Series, 3/2014

Growth, Inequality and Poverty Reduction in Africa, by Francisco H.G. Ferreira, Milan, FEEM Press, Economy and Society Series, 1/2015

Coalitions and Networks, by Carlo Carraro (ed.), Milan, FEEM Press, Climate Change and Sustainable Development Series, 2/2015

The Coalition Theory Network (CTN) is an association of eight high-level scientific and academic institutions whose aim is the advancement and diffusion of research in the area of network theory and coalition formation. The history of CTN began in 1995, when FEEM joined CORE – University of Louvain la Neuve in organizing a workshop on coalition formation and environmental games, with focus on the analysis of the process of international climate negotiations.

The success of the workshop induced the organisers to widen the focus of the following meetings to the burgeoning applications of coalition and network theory, and to undertake the formal creation of CTN. The yearly meetings have continued for 20 years, hosted in turn by the partner institutions, among which those that have joined in the meantime (GREQAM - University of Aix Marseilles, and CES – University Paris I in 1999, MOVE – Universitat Autònoma de Barcelona in 2000, Maastricht University and Vanderbilt University in 2006, and CSDSI – New Economic School in 2014).

The CTN has progressively become a reference point for the study of network and coalition formation and their applications. This volume represents a tribute to research developed by the Coalition Theory Network and presented at the CTN annual workshops, on the occasion of the 20th anniversary of its foundation.

Carlo Carraro is Professor of Environmental Economics and Econometrics at Ca' Foscari University of Venice, and Director of the Climate Change and Sustainable Development Programme of the Fondazione Eni Enrico Mattei. He is Vice-Chair of the Working Group III and Member of the Bureau of the Intergovernmental Panel on Climate Change (IPCC) and Co-Chair of the Green Growth Knowledge Platforms' Advisory Board. He is also Director of the International Center for Climate Governance. Professor Carraro has written more than 200 papers and 30 books on the international coordination of monetary policy, coalitions and group formation in economic systems, international negotiations and the formation of international environmental agreements, climate change modelling and policy.

coalition theory network