

# Essays on the Evolution of Preferences and Network Interactions

Thesis submitted

in partial fulfilment of the requirements for the

JOINT DOCTORAL DEGREE IN  
ECONOMICS AND APPLIED MATHEMATICS

by

Olena Orlova

Universität Bielefeld, Université Paris 1 Panthéon-Sorbonne

July 2021

EXAMINATION COMMITTEE:

SUPERVISORS:

Prof. Dr. Frank Riedel, Universität Bielefeld

Prof. Dr. Agnieszka Rusinowska, Université Paris 1 Panthéon-Sorbonne

Prof. Dr. Christoph Kuzmics, Universität Graz

EXTERNAL REVIEWERS:

Prof. Dr. Tim Hellmann, University of Southampton

Prof. Dr. Penélope Hernández, Universitat de València

Universität Bielefeld  
Fakultät für Wirtschaftswissenschaften  
Universitätsstraße 25  
33615 Bielefeld  
Germany

---

Université Paris 1 Panthéon-Sorbonne  
Centre d'Économie de la Sorbonne  
106-112 Boulevard de l'Hôpital  
75013 Paris  
France

## Abstract

This doctoral thesis explores the theme of preference evolution in the course of social interactions, as well as the interplay between individuals' preferences and a social environment and the impact of both on social outcomes. It consists of three parts.

The first part studies the evolution of social preferences such as altruism, selfishness or reciprocity when individuals are repeatedly involved in one-shot bilateral interactions modeled by the prisoner's dilemma game. Individuals are heterogeneous not only in their social preferences (subjective preferences over the outcomes of the game) but also in their ability to observe their opponents' preferences (so-called cognitive intelligence). Coevolution of social preferences and cognitive intelligence is studied within both static and dynamic frameworks.

The second part delves deeper into the composition of individuals' preferences. It disentangles the idiosyncratic and the interactional preference components and studies their interplay in the framework where interactions take place on a fixed network. It appears that heterogeneity in idiosyncratic preferences changes equilibrium outcomes in a non-trivial fashion: some equilibria disappear and qualitatively new ones appear instead. A particular outcome, in which everyone's idiosyncratic preferences are satisfied, is a unique efficient outcome in many games on networks, but it is not always an equilibrium.

The third part further develops the proposed framework and considers how heterogeneous preferences can influence the formation of an interactional structure (a network). In this model individuals are allowed to choose their interaction partners simultaneously with their action choice in each interaction. Despite the symmetry and simplicity of the setting (binary action choice and two types of idiosyncratic preferences), quite irregular network structures can arise in equilibrium. This finding suggests that heterogeneity in action preferences may already explain a large part of observed irregularity in endogenously formed networks.

## Resumé

Cette thèse de doctorat explore le thème de l'évolution des préférences au cours des interactions sociales, ainsi que réciproque entre les préférences des individus et un environnement social et l'impact des deux sur les résultats sociaux. Il se compose de trois parties.

La première partie étudie l'évolution des préférences sociales telles que l'altruisme, l'égoïsme ou la réciprocité lorsque les individus sont impliqués de manière répétée dans des interactions bilatérales ponctuelles modélisées par le jeu du dilemme du prisonnier. Les individus sont hétérogènes non seulement dans leurs préférences sociales (préférences subjectives sur les résultats du jeu) mais aussi dans leur capacité à observer les préférences de leurs adversaires (ce qu'on appelle l'intelligence cognitive). La coévolution des préférences sociales et de l'intelligence cognitive est étudiée dans des cadres à la fois statiques et dynamiques.

La deuxième partie approfondit la composition des préférences des individus. Il démêle les composantes de préférence idiosyncratique et interactionnelle et étudie leur réciproque dans le cadre où les interactions ont lieu sur un réseau fixe. Il apparaît que l'hétérogénéité des préférences idiosyncratiques modifie les résultats de l'équilibre de manière non triviale: certains équilibres disparaissent et de nouveaux, qualitativement, apparaissent à la place. Un résultat particulier, dans lequel les préférences idiosyncratiques de chacun sont satisfaites, est un résultat efficace unique dans de nombreux jeux sur les réseaux, mais ce n'est pas toujours un équilibre.

La troisième partie développe davantage le cadre proposé et considère comment des préférences hétérogènes peuvent influencer la formation d'une structure interactionnelle (un réseau). Dans ce modèle, les individus sont autorisés à choisir leurs partenaires d'interaction simultanément avec leur choix d'action dans chaque interaction. Malgré la symétrie et la simplicité du cadre (choix d'action binaire et deux types de préférences idiosyncratiques), des structures de réseau assez irrégulières peuvent apparaître en équilibre. Cette découverte suggère que l'hétérogénéité des préférences d'action peut déjà expliquer une grande partie de l'irrégularité observée dans les réseaux formés de manière endogène.

# Contents

Acknowledgements . . . . .	3
Introduction . . . . .	5
<b>1 Evolution of Social Preferences and Cognitive Intelligence</b>	<b>8</b>
1.1 Introduction . . . . .	8
1.2 The model . . . . .	12
1.2.1 Fitness game, types and configurations . . . . .	12
1.2.2 Type game and stability concepts . . . . .	14
1.3 Static framework . . . . .	16
1.4 Dynamic framework . . . . .	20
1.5 Extension: Public preference and deception . . . . .	22
1.6 Conclusions . . . . .	23
Appendix to Chapter 1 . . . . .	24
<b>2 Idiosyncratic Preferences in Games on Networks</b>	<b>29</b>
2.1 Introduction . . . . .	29
2.2 The model . . . . .	33
2.2.1 Games and idiosyncratic action preferences . . . . .	33
2.2.2 Equilibrium concept . . . . .	35
2.3 Best response functions . . . . .	36
2.3.1 Coordination games . . . . .	37
2.3.2 Anti-coordination games . . . . .	38
2.3.3 Dominant action games . . . . .	40
2.3.4 Companion/opponent requirement . . . . .	41
2.4 Equilibrium analysis . . . . .	41
2.4.1 Coordination games . . . . .	42
2.4.2 Anti-coordination games . . . . .	44
2.4.3 Dominant action games . . . . .	46
2.4.4 Efficiency of equilibria . . . . .	47
2.5 Discussion . . . . .	53

2.6	Conclusions . . . . .	55
	Appendix to Chapter 2 . . . . .	56
<b>3</b>	<b>Network Games with Heterogeneous Players</b>	<b>65</b>
3.1	Introduction . . . . .	65
3.2	The model . . . . .	67
	3.2.1 The game . . . . .	67
	3.2.2 Equilibrium concept and some graph theory notions . . . . .	69
3.3	Equilibrium analysis . . . . .	70
	3.3.1 Preliminaries . . . . .	70
	3.3.2 Classes of equilibria . . . . .	72
3.4	Discussion and conclusions . . . . .	79
	Appendix to Chapter 3 . . . . .	82
	<b>Bibliography</b>	<b>99</b>

# Acknowledgements

First of all, I would like to express my deep gratitude to my thesis supervisors Christoph Kuzmics, Frank Riedel, and especially Agnieszka Rusinowska, for their constant and all-around support during my PhD years. Agnieszka was a true "Doktormutter" for me in all respects and, among other things, taught me the important skill of disseminating my research effectively and in various ways. Without her immense support and continuous encouragement, this work would have never taken shape. Conversations with Christoph have always been one of the most powerful sources of scientific inspiration for me. Whenever I needed a refill for my enthusiasm or the courage to start a new scientific adventure, I knew where to find it. Last but not least, Frank's support and guidance were indispensable in the last stage of my PhD, also in terms of such an essential component as research funding. It is to him I owe, in particular, all the time I got to perfect my last article. I really appreciate the excellent scientific advising I was given by all of my supervisors and their generosity in sharing it.

I am also very grateful to Tim Hellmann and Penélope Hernández, who kindly agreed to join the examination committee and whose valuable feedback on earlier stages helped to significantly improve my work.

The list of people who found time and patience to listen to and comment on my research presentations during numerous conferences, seminars and personal talks is so long that I could not hope to name everyone here. I would like to thank Francis Bloch, Mira Frick, Dunia López-Pintado, François Maniquet, Noemí Navarro, Sudipta Sarangi, Joel Sobel, Marina Uzunova, Fernando Vega-Redondo, all the members of the Center for Mathematical Economics and the Faculty of Economics at Bielefeld University and all the participants of the research group "Networks and Games" at Centre d'Économie de la Sorbonne, as well as two anonymous referees and the advising editor of *Games and Economic Behavior* for their valuable suggestions and generous comments.

Several conference discussions had a particularly significant impact on my research, and I would like to thank separately to all the participant of these discussions at the summer school on network theory CIGNE-2017, the 13<sup>th</sup> and 14<sup>th</sup> SING European meetings on game theory, LEG2019: Learning, evolution and games conference and the invited seminar talk at the University of Maastricht. I also deeply appreciate our interdisciplinary discussions with Martin Eriksen and his technical assistance in my first research article.

My warm thanks go to Diana Grieswald, Bettina Robson, Liliana de Freitas and Nathalie Louni for their empathy and being such a great help to me with moving and adjusting to the new universities. My sincere thanks also to all the students in Paris and Bielefeld who made my university life much brighter. Lalaina, Zainab, Torben, Jinglin, Cuong, Léa, Marine, Natasha, Carla, Alexis, Martin, Herman, Simon, Philipp, Marieke, Bulgan, Daniel, Ilya,

Cynda, Mustapha, Okay, Alessandro, Sevak and all the others, thank you for being around.

I want to thank my son for being patient and understanding (most of the time) and to my dear sisters and my wonderful friends Olga, Farid and Shaza for taking over my parental responsibilities during my research trips. I am deeply indebted to my parents, who supported me devotedly during all these years regardless of the distance between us. This is truly invaluable.

I acknowledge financial support of the European Commission in the framework of the European Doctorate in Economics - Erasmus Mundus (EDEEM), the French Ministry of Europe and Foreign Affairs in the framework of the Eiffel Excellence Scholarship Program and the Franco-German University (DFH-UFA). I would also like to thank the French National Agency for Research (ANR), Bielefeld Graduate School for Economics and Management (BiGSEM), École doctorale d'Économie Panthéon-Sorbonne (ED 465) and the University of Graz for funding my conference participation and research stays.

The usage of "we" in my articles is a tribute to my scientific supervisors. All remaining mistakes are only mine.



# Introduction

Individuals' preferences lie in the foundations of microeconomics and enter into assumptions of most macroeconomic models. Most of the decision problems that individuals are confronted with, however, imply dependence on decisions of other individuals, each of whom has their own preferences. These preferences might be closely aligned, partially aligned or not at all, making the involved individuals strategic players in a particular game.

In this thesis I construct several different game-theoretical models that allow me to study various aspects of the specific interplay between individuals' preferences, interactional structures in which these individuals are involved, and the nature of their interactions (the rules of a particular game).

Interaction rules are assumed exogenous and fixed. I consider a wide variety of two-player games with binary action choice (2x2 games) that incentivize players either to coordinate their actions, to anti-coordinate, or to follow a particular dominant action. I study how these rules influence the evolution of individuals' preferences (Chapter 1) and the formation of interactional structures (Chapter 3).

I also explore the relationship between preferences and interactional structures. In the first two chapters an interactional structure is fixed: in Chapter 2 it is an arbitrary fixed network and in Chapter 1 it is a complete network, in which each individual can be randomly matched for an interaction with any other individual. In Chapter 3, however, the interactional structure is a result of a network game, in which each player receives the sum of her payoffs in 2x2 games with everyone with whom she establishes connections. Chapter 3 thus allows to study the impact of heterogeneous preferences on the formation of an interactional structure.

Finally, I study the significance of all three components – individuals' idiosyncratic preferences, an interactional structure and interaction rules – for equilibrium outcomes of the game. In Chapter 2, I compare equilibria of different games on arbitrary fixed networks when players have heterogeneous action preferences. In Chapter 3, where the interactional structure is endogenous, the notion of an equilibrium outcome includes, apart from an action profile, the associated endogenously formed network. It allows for comparison of equilibrium action profiles in the case of exogenous and in the case of endogenous network structures.

Below I briefly introduce the models presented in the following chapters and the main results they produce.

**Chapter 1. Evolution of social preferences and cognitive intelligence.** The first chapter contributes to the literature on the evolution of preferences (see Güth and Yaari (1992), Samuelson (2001), Robson and Samuelson (2011)). It develops a model of preference evolution in which preferences and their observability coevolve, thus allowing to consider situations different from the extreme cases of complete observability/unobservability of pref-

erences (see Dekel et al. (2007) and Herold and Kuzmics (2009) for the analysis of some partial observability scenarios).

In my model individuals from an infinite population are repeatedly randomly matched to play the prisoner’s dilemma game. Each individual is characterized by a social preference (a subjective preference over outcomes of the game) and a so-called cognitive intelligence level that determines her ability to observe her opponents’ preferences. With regard to the latter, all players are subdivided into two types – naïve, unaware of their opponents’ preferences, and intelligent, who have complete information about their opponents. Being intelligent is costly, which excludes the possibility of a ”secret handshake” like in Robson (1990).

I completely characterize neutrally stable configurations of players with the same cognitive intelligence level, and find that my results are consistent with those of other related studies (in particular, Dekel et al. (2007) and Heller and Mohlin (2019)). In stable configurations with naïve players (which correspond to complete unobservability of preferences) everyone defects, playing a Nash equilibrium of the fitness game, while in stable configurations with intelligent players (corresponding to the complete observability case) everyone cooperates, which is the only efficient outcome.

I also derive necessary conditions for neutral stability of configurations of players with heterogeneous cognitive intelligence levels. These conditions suggest that, if such configurations exist, they include quite a limited set of social preferences. Simulations of the replicator dynamics reveal that some cognitively heterogeneous configurations indeed show (weaker) stability properties, following a cyclical pattern.

**Chapter 2. Idiosyncratic preferences in games on networks.** The second chapter contributes to the literature on games played on networks (see Galeotti et al. (2010), Jackson and Zenou (2014), Bramoullé and Kranton (2016)) between players with heterogeneous action preferences. It extends the framework proposed in Hernández et al. (2013) and studies the interplay between individuals’ idiosyncratic preferences and interactional incentives implied by the rules of a particular game.

I consider a broad class of 2x2 games, including games with strategic complements and with strategic substitutes, played between each pair of connected players on a fixed network. Everyone’s payoff is the sum of her payoffs with each of her network partners. Interactional incentives are the same for all players, while their action preferences differ. I investigate the equilibrium outcomes for different games and compare them with the outcomes under homogeneous players’ preferences: generically, there is no inclusion of these equilibrium sets in either direction.

I show that existence of two sufficiently segregated (interconnected) subsets of players, irrespective of their idiosyncratic preferences, is necessary and sufficient for existence of a het-

erogeneous action equilibrium in coordination (anti-coordination) games. In dominant action games, the action profile in which every player chooses her preferred action is a unique equilibrium. Such an equilibrium might also exist in coordination and anti-coordination games if groups of players with identical preferences satisfy corresponding connectivity conditions.

Since experimental studies of equilibrium selection in games on networks (see e.g. Char-ness et al. (2014)) show that under complete information socially efficient equilibria are typically implemented, I attempt to perform some efficiency analysis of the equilibria. I find that in many games, the equilibria in which all players choose their preferred actions are unique efficient equilibria, and in yet more games they are Pareto efficient.

**Chapter 3. Network games with heterogeneous players.** The last chapter extends the analysis of the previous one to the setting in which players, simultaneously with their action choice, choose their interaction partners. This work contributes to the literature on network formation games with a simultaneous action choice (see Jackson and Watts (2002), Bramoullé et al. (2004), Goyal and Vega-Redondo (2005), Baetz (2015), Hiller (2017)), adding a new dimension of heterogeneity between players – their idiosyncratic action preferences.

In my model link formation is two-sided, that is, creation of a link takes place if and only if a pair of players makes mutual link proposals. The linking cost is strictly positive. As before, a player’s payoff is the sum of her payoffs with every network partner. I characterize equilibria for different games, varying not only the strength of action preferences but also the linking cost. Compared to the previous chapter, endogenizing the network structure allows for much more precise equilibrium characterizations.

Goyal et al. (2021) distinguish two types of equilibria that arise in a similar setting in a coordination game: either all players coordinate on the same action in a complete network, or they all choose their preferred actions and form disjoint action cliques. I find that quite irregular network structures with action variety but not perfect preference satisfaction can also appear in equilibrium. Interestingly, such irregular equilibrium structures exist both for coordination and anti-coordination games, and their existence is robust to equilibrium refinements in which all kinds of bilateral deviations are admissible.

It is worth noting that the most irregular equilibrium structures can exist only for intermediate linking cost values. Nevertheless, the very possibility of their existence implies that even players’ preference heterogeneity alone can explain a large part of observed irregularity in endogenous network structures.

# Chapter 1

## Evolution of Social Preferences and Cognitive Intelligence

**Abstract:** This paper contributes to the literature on the evolution of preferences by developing a model with endogenous observability. Individuals are randomly and repeatedly matched to play the prisoner's dilemma. Each individual is characterized by a subjective preference over outcomes of the game (a social preference) and a level of cognitive intelligence, which determines her ability to observe her opponent's preference. We examine coevolution of social preferences and cognitive intelligence both within static and dynamic frameworks.

**JEL codes:** C72, C73, D83, D91.

**Keywords:** *evolution of preferences; indirect evolutionary approach; theory of mind; cognitive sophistication.*

### 1.1 Introduction

Since most of the mainstream economics models take preferences as given, the question of which preferences are more viable in a given environment and thus more plausible to be assumed comes naturally. Shaping of social preferences, such as altruism, spite or reciprocity, seems to be heavily influenced by the outcomes of social interactions. For this reason, evolutionary game theory approach seems to be especially well suited for studying evolution of this type of preferences.<sup>1</sup>

Since its introduction by Güth and Yaari (1992), the central tool for studying the evolution of preferences has been the indirect evolutionary approach, according to which preferences motivate behavior, behavior determines success, and success in its turn shapes future

---

<sup>1</sup>A quite comprehensive study of the evolutionary foundations of preferences can be found in Robson and Samuelson (2011).

preferences. More specifically, the causal loop can be described as follows. Individuals are characterized by their preferences rather than simply actions, and these preferences induce behavior (actions) through a choice procedure.<sup>2</sup> Some social behaviors prove to be objectively more successful, more beneficial for those who exhibit them than other behaviors. During social interactions, individuals observe the outcomes and revise their behaviors, sometimes through trial and error, to adapt to the existing social environment. As a result, the behaviors that are more successful are reproduced more often, and social preferences that survive in society are those that are able to induce them.

This new approach received an enthusiastic welcome in 1990s and early 2000s and resulted in a large literature, which used it to explain the existence of preferences different from self-interest.<sup>3</sup> However, most of this literature shares two important limitations: it implies at least partial observability, and narrows down the class of admissible preferences to a specific set.<sup>4</sup> Under such conditions many "non-standard" preferences appear to be stable. On the other hand, none of such preferences has proven stable under no observability. Many studies confirm the same result: when preferences are not observable, only Nash outcomes are stable, which for the case of the prisoner's dilemma means selfish defection.<sup>5</sup>

There are several works departing from the polar cases of perfect and no observability and considering intermediate degrees of observability. Dekel et al. (2007) analyze almost perfect and almost no observability cases as a robustness check for their results. Herold and Kuzmics (2009) compare the outcomes of preference evolution under perfect and almost perfect observability. Although these models do not cover the whole range of intermediate cases, their conclusions convincingly confirm that varying the degree to which preferences are observed results in different outcomes of preference evolution.

There definitely exists a psychological rationale behind the assumption of observability of preferences.<sup>6</sup> In particular, emotions, being mostly out of our conscious control, often may serve as sensitive indicators of our preferences. At the same time, an accurate reconstruction of preferences from observed emotions is not always possible. Moreover, there is room for mimicry (we will discuss it below in more detail), and in such a case reconstruction of preferences from observed signals would be even misleading. Güth (1995) argues that the

---

<sup>2</sup>Utility maximization is one of the most typical choice procedures assumed in economics; it is also used in most of the literature applying the indirect evolutionary approach.

<sup>3</sup>Among others, see Bester and Güth (1998) for the model explaining altruistic preferences, Bolle (2000) for extending their results to spiteful preferences, Güth and Napel (2006) for analysis of inequality aversion in a multiple games environment, Sethi and Somanthan (2001) for investigating viability of reciprocal preferences conditional on the opponent's preference type, Kockesen et al. (2000) for evolution of a more general form of interdependent preferences.

<sup>4</sup>This critique was originally expressed in Samuelson (2001).

<sup>5</sup>See, for example, Ok and Vega-Redondo (2001), Ely and Yilankaya (2001), Dekel et al. (2007).

<sup>6</sup>Frank (1988) discusses the technological basis for acquiring information about others' preferences.

assumption that people never have information about the preferences of others would be as strong and unrealistic as the assumption of perfect observability of preferences, implying that exactly these intermediate cases are of particular interest.

Robson and Samuelson (2011) further develop this argument: "... we can observe preferences because people give signals – a tightening of the lips or flash of the eyes – that provide clues as to their feelings. However, the emission of such signals and their correlation with the attendant emotions are themselves the product of evolution. [...] the indirect evolutionary approach will remain incomplete until the evolution of preferences, the evolution of signals about preferences, and the evolution of reactions to these signals, are all analyzed within the model." Thus comes a key conclusion: observability of preferences can (and should) be endogenized within the evolutionary framework.

In this respect, it is reasonable to discuss two important issues. First of them, as have been already mentioned, is the possibility of mimicry.<sup>7</sup> Once individuals develop the ability to signal their preferences, there might always be an incentive to give a misleading signal and deceive the opponent about own true preference.<sup>8</sup> There is a clear evidence that some people utilize these possibilities better than others do. On the one hand, they might be more capable of imitation, of sending misleading signals about their social preferences, for example, pretending to be more cooperative than they actually are. On the other hand, they might be good at seeing through the others, regardless of who their opponents pretend to be and what they claim to prefer. Heller and Mohlin (2019) consider these two abilities together and call it the level of *cognitive sophistication* of an individual.<sup>9</sup> They take the set of all natural numbers to represent possible levels of cognitive sophistication and consider a strong form of deception: whenever two cognitively different individuals meet, the more sophisticated one can perfectly observe the preference of the other and, at the same time, can choose whatever she wants the deceived opponent to believe about her preference.

In this paper, we propose a model that builds on the idea of Heller and Mohlin (2019) but differs from it in several important respects. We depart from the assumption of strong deception and, furthermore, separate the imitational and observational dimensions of cognitive ability. Even though these dimensions seem to be correlated, they constitute different abilities and might have evolved as interdependent but separate traits. Since we are concerned with endogenous observability of preferences, our analysis focuses on the ability to observe preferences of others – the feature we name *cognitive intelligence*. Preference mimicry is the next evolutionary step, which undoubtedly also matters for the evolution of observability.

---

<sup>7</sup>There is ample evidence of mimicry from the animal world. Thorough studies on this topic include Wickler (1968), Ruxton et al. (2004) and Maynard Smith and Harper (2007).

<sup>8</sup>This phenomenon is discussed in Samuelson (2001) and, in more detail, in Robson and Samuelson (2011).

<sup>9</sup>A related strand of literature is the one on the theory of mind. See Robalino and Robson (2012) for the detailed summary of economic and game theoretical models.

We consider it as an extension to our model. Another difference consists in the fact that we consider a somewhat simplified setting with regard to the range of cognitive intelligence levels. We assume that an individual can be either *naïve*, that is, not able to observe the preference of her opponent at all, or *intelligent*, able to observe her opponent’s preference perfectly. Such a simplification might appear quite strong at first sight; however, as the results of Heller and Mohlin (2019) suggest, what matters is just being more sophisticated than the opponent.

The second issue worth mentioning with regard to the evolution of observability is the cost of developing the ability to produce and interpret the signals, or following the terminology of Heller and Mohlin (2019), the *cognitive cost*. It is obvious that this cost must not be very high to allow some individuals to increase their level of cognition. On the other hand, if the cognitive cost is zero, we might observe something similar to the phenomenon of “secret handshake” described in Robson (1990), demonstrating instability of any inefficient equilibrium. Wiseman and Yilankaya (2001), building on this idea, used simulations to illustrate a possible 3-stage cycle of the evolution of preferences, in which individuals of different levels of cognition temporarily co-exist. In this paper, we are primarily interested in the cases when the cognitive cost is strictly positive but sufficiently small to allow for increase in cognition.

We consider an infinite population of individuals who are repeatedly and randomly matched to play a fitness game – the prisoner’s dilemma.<sup>10</sup> Each individual is endowed with a subjective preference over outcomes of the game, which does not necessarily correspond to her objective fitness payoffs associated with these outcomes. In addition, each individual possesses a cognitive characteristic (we differentiate between naïve and intelligent players), which determines the information the player has about her opponent’s preference. This subjective preference together with the information about the opponent dictate behavior of the player. Once the game is played, all the parties receive fitness payoffs accordingly, and the composition of the population evolves: those types that earned higher fitness grow at the expense of those that earned lower. Since the notion of a *type* in our model combines both a preference component and a cognitive component, it enables us to study coevolution of preferences and observability.

The rest of the paper is organized as follows. Section 1.2 introduces the formal model, as well as the solution concepts we use. Sections 1.3 and 1.4 present the analysis of the model within static and dynamic frameworks respectively. Section 1.5 outlines some directions for extending the model, and section 1.6 concludes. Appendix contains the proofs of the results

---

<sup>10</sup>Since it has been shown that both finiteness of a population and assortative matching favor evolution of non-selfish preferences, we avoid these assumptions in order not to mix their effects with the effect of endogenizing observability. See, for example, Alger and Weibull (2013) for assortative matching and Schaffer (1988) for finite populations.

for the static solution framework.

## 1.2 The model

### 1.2.1 Fitness game, types and configurations

A *fitness game*  $G$  is a two-player symmetric game with a finite set of actions  $A$  and a fitness payoff function  $\pi : A \times A \rightarrow \mathbb{R}$ . The payoff function  $\pi$  extends in the standard way to mixed actions:  $\pi(\sigma, \sigma')$  denotes the expected fitness payoff to a player playing a (mixed) action  $\sigma \in \Delta(A)$  against an opponent playing  $\sigma' \in \Delta(A)$ .

We restrict our attention to the prisoner's dilemma as the underlying fitness game. Therefore, the set of actions is  $A = \{C, D\}$  with  $C$  standing for cooperation and  $D$  for defection, and the fitness payoffs are the following:

	$C$	$D$
$C$	$1, 1$	$-x, y$
$D$	$y, -x$	$0, 0$

where  $x > 0$ ,  $y > 1$  and  $y - x < 2$ .<sup>11</sup>

Every individual has a subjective preference over outcomes in  $G$ , represented by a subjective von Neumann-Morgenstern utility function  $u : A \times A \rightarrow \mathbb{R}$ , which, generically, is different from the objective fitness payoff function  $\pi$ . In order not to restrict the set of conceivable preferences that can arise in the course of evolution, we make no further specific assumptions about  $u$ , and denote by  $U$  the set of all possible subjective utility functions. None of the individuals knows the true payoffs of the fitness game  $G$ ; they condition their behavior solely on their subjective preferences over outcome profiles. However, once the game has been played, all players receive their fitness payoffs according to  $\pi$ .

A *type* of a player is a pair, consisting of a subjective utility function  $u$  and a cognitive intelligence level  $n$ :  $\theta = (u, n) \in \Theta = U \times \{0, 1\}$ . We also use the notation  $u_\theta$  and  $n_\theta$  to refer to the subjective preferences and the cognitive level of type  $\theta$ . If the cognitive level equals zero, we call such a type *naïve* and assume that players of this type cannot observe their opponent's type at all – neither her preferences, nor her cognitive level. If the cognitive level equals one, we call such a type *cognitively intelligent* (or simply *intelligent*) and assume that players of this type can observe the type of their opponent perfectly. We also make a standard assumption that all players know the general type distribution of the

---

<sup>11</sup>Without loss of generality, we can normalize to 1 the difference between payoffs for mutual cooperation and for mutual defection. The last inequality,  $y - x < 2$ , guarantees that mutual cooperation is an efficient outcome.



population. This assumption can be justified either by prolonged individual learning or by public availability of the outcomes of all interactions.

We assume that, given their types and beliefs about types of their opponents, players behave rationally. Given a type distribution  $\mu \in \Delta(\Theta)$  with finite support  $C(\mu)$ , we can define a *behavior policy*  $b : C(\mu) \times C(\mu) \rightarrow A$  that describes optimal behavior for each player in each possible match.<sup>12</sup> A particular behavior policy  $b$  implies a Bayesian-Nash equilibrium of the game in subjective preferences, that is, the game in which payoffs to the players are given by their subjective utility functions. In our model, an intelligent player observes perfectly her current opponent's type and thus can always best-reply to her opponent's action. A naïve player, however, has to consider the general type distribution of the population and, being unable to distinguish one opponent from another, plays the same action in every match – the best-reply to the average play towards her.

We assume that a pure equilibrium is played whenever one exists. If a player is indifferent between two actions, we let exogenous factors choose the *focal action*  $f(C, D) \in \{C, D\}$  that will be played.<sup>13</sup> The behavior policy for each pair of types  $\theta, \theta' \in C(\mu)$  is given by

$$b(\theta, \theta') = b_\theta(\theta') = \begin{cases} f(BR_\theta(b_{\theta'}(\theta))) & \text{if } n_\theta = 1, \\ f(BR_\theta(b.(\theta); \mu)) & \text{if } n_\theta = 0, \end{cases} \quad (1.1)$$

where  $BR_\theta(\sigma)$  denotes the set of best-replies to  $\sigma$  given preferences  $u_\theta$ , and  $BR_\theta(b.(\theta); \mu) = BR_\theta(\sum_{\theta_i \in C(\mu)} \mu(\theta_i) b_{\theta_i}(\theta))$ . We interpret  $b_\theta(\theta')$  as the strategy of a player of type  $\theta$  matched to play the game with a player of type  $\theta'$ .

Following Dekel et al. (2007) and Heller and Mohlin (2019), we now define a configuration:

**Definition 1.** A *configuration* is a pair  $(\mu, b)$ , where  $\mu \in \Delta(\Theta)$  is a type distribution with finite support  $C(\mu)$ , and  $b : C(\mu) \times C(\mu) \rightarrow A$  is a behavior policy such that (1.1) holds for every pair of types  $\theta, \theta' \in C(\mu)$ .

Let us emphasize that a configuration, that is, a type distribution *together with* equilibrium behavior, is a reasonable basic unit for studying stability, since, in general multiple equilibria might exist for the same type distribution.

As in the standard evolutionary game theory framework, all individuals are randomly and repeatedly matched to play the fitness game  $G$ . Their behavior is determined by their subjective preferences and the information they possess about their opponents (which, in its turn, is determined by their cognitive intelligence levels). Evolutionary logic is reflected in the dynamics of the population: the change of each type's share is proportional to the

---

<sup>12</sup>The term *behavior policy* is borrowed from Heller and Mohlin (2019).

<sup>13</sup>In the general case, a *focality function*  $f : 2^A \rightarrow A$  is such that  $f(S) \in S$  for any  $S \in 2^A$ .

relative fitness earned by this type. We seek to characterize stable configurations, in which the shares and behavior of each type remain constant despite small perturbations to the type distribution. The precise definitions of what we mean by stability in this paper are presented in the next subsection.

## 1.2.2 Type game and stability concepts

Every configuration  $(\mu, b)$  induces a *type game*  $\Gamma_{\mu, b}$  – a two-player symmetric game in which strategies are different types  $\theta \in C(\mu)$  and payoffs are their fitness payoffs when matched with each other. That is, for every type  $\theta$  when matched with  $\theta'$ , its payoff in  $\Gamma_{\mu, b}$  equals  $\pi(b_\theta(\theta'), b_{\theta'}(\theta)) - k_\theta$ , where  $k_\theta$  is the *cognitive cost* of type  $\theta$ . We let  $k_\theta = n_\theta \cdot k$  with  $k > 0$ , that is, the cost is zero for naïve types and strictly positive for intelligent types.

For every type  $\theta$ , its expected fitness payoff in  $\Gamma_{\mu, b}$  is, consequently,

$$\Pi_\theta(\mu, b) = \sum_{\theta' \in C(\mu)} \mu(\theta') \cdot \pi(b_\theta(\theta'), b_{\theta'}(\theta)) - k_\theta,$$

that is, its expected material payoff net the cognitive cost.

Having defined a type game, we can extend the standard notions of an evolutionary stable strategy (ESS) and a neutrally stable strategy (NSS) from strategies to configurations.<sup>14</sup> Let us remind that a strategy  $\sigma$  is called *evolutionary stable* if for every mutant strategy  $\sigma'$  there exists an invasion barrier  $\bar{\varepsilon} > 0$  such that for every positive  $\varepsilon < \bar{\varepsilon}$  it holds that  $\pi(\sigma', (1 - \varepsilon)\sigma + \varepsilon\sigma') < \pi(\sigma, (1 - \varepsilon)\sigma + \varepsilon\sigma')$ ; that is, whatever is the mutant strategy it cannot invade the population (as it earns less fitness) as long as its share is below the invasion barrier. A strategy is called *neutrally stable* if the above inequality is not strict.

Before defining evolutionary and neutrally stable configurations, however, we have to define a focal configuration, which specifies some relevant post-entry equilibria.<sup>15</sup>

**Definition 2.** A configuration  $(\tilde{\mu}, \tilde{b})$  is *focal* with respect to  $(\mu, b)$  if  $C(\mu) \subseteq C(\tilde{\mu})$  and for every  $\theta, \theta' \in C(\mu)$  it holds that  $\tilde{b}_\theta(\theta') = b_\theta(\theta')$ .

To put it another way, a post-entry configuration is focal if incumbent behavior is unchanged after the entry of mutants.

Now we can introduce the definitions of static stability of a configuration.

**Definition 3.** A configuration  $(\mu, b)$  is *evolutionary stable* if for every mutant type  $\theta' \in \Theta$  there exists an invasion barrier  $\bar{\varepsilon} > 0$  such that for every positive  $\varepsilon < \bar{\varepsilon}$ :

<sup>14</sup>The notion of an evolutionary stable strategy was introduced by Maynard Smith; see Maynard Smith and Price (1973) and Maynard Smith (1982).

<sup>15</sup>We borrow this idea from Dekel et al. (2007). A similar definition appears in Heller and Mohlin (2019).

- (i)  $\exists(\tilde{\mu}, \tilde{b})$  such that  $(\tilde{\mu}, \tilde{b})$  is focal w.r.t.  $(\mu, b)$ ;
- (ii) for every  $(\tilde{\mu}, \tilde{b})$  that satisfies (i),  $\Pi_{\theta'}(\tilde{\mu}, \tilde{b}) < \Pi_{\mu}(\tilde{\mu}, \tilde{b})$ .

A configuration is *neutrally stable* if the last equality is not strict.

In other words, a configuration  $(\mu, b)$  is evolutionary (neutrally) stable if for every sufficiently small-scale monomorphic mutation (i) there exists at least one focal post-entry configuration, and (ii) for every focal post-entry configuration  $(\tilde{\mu}, \tilde{b})$ ,  $\mu$  is an evolutionary (neutrally) stable strategy in the type game  $\Gamma_{\tilde{\mu}, \tilde{b}}$ .<sup>16,17</sup>

If there exists a mutation which necessarily forces incumbents to change their behavior among themselves, we regard the original configuration as unstable (hence the first condition of the definition). Moreover, every mutant in every focal post-entry configuration, should earn less in terms of fitness – or at least, in the case of neutral stability, not more – than the incumbent configuration as a group (hence the second condition of the definition).

It is worth noting that the stronger of these two stability notions, evolutionary stability, is not useful in our setting. We can always construct a new mutant type with the same cognitive intelligence level as one of the incumbent types, but slightly different preferences that nevertheless induce the same behavior. Such mutants can earn the same fitness as incumbents, and thus evolutionary stable configurations never exist. In the analysis that follows we use the weaker static stability notion, neutral stability. Below we will use the abbreviation NSC for a neutrally stable configuration.

We still have to make several further assumptions concerning possible mutations and the speed of evolution. First, we make the standard for the indirect evolutionary approach assumption that adjustment of behavior happens infinitely faster than adjustment of the type distribution. Suppose, the fraction  $\varepsilon$  of the population in the original configuration  $(\mu, b)$  is replaced by mutants – individuals of some type  $\theta' \in \Theta$ . The post-entry type distribution is then  $\tilde{\mu} = (1-\varepsilon)\mu + \varepsilon\theta'$ . The above assumptions means that the post-entry behavior transforms very quickly into a new  $\tilde{b}$ , well before fitness differences between types start changing the

---

<sup>16</sup>It is important to note that our definition of an evolutionary (neutrally) stable configuration slightly differs from corresponding definitions in other studies. In particular, Heller and Mohlin (2019) do not require existence of focal equilibria, it suffices that condition (ii) of the definition holds. Dekel et al. (2007) substitute the requirement of existence of a focal equilibrium with existence of at least a "nearby" equilibrium, and expect condition (ii) to hold for all such "nearby" equilibria. Thus, in this sense our definition is stronger than the above ones. On the other hand, however, Heller and Mohlin (2019) permit polymorphic mutant groups, while we consider only monomorphic mutants.

<sup>17</sup>Oechssler and Riedel (2002) propose an alternative static evolutionary stability concept, *evolutionary robustness*, for models in which, like in ours, the strategy space is continuous. It is a stronger stability concept and it has been shown to guarantee dynamic stability for replicator dynamics in doubly symmetric games. So far, we intend not to restrict the set of stable configurations too much and will thus consider stability in a broader sense.

distribution  $\tilde{\mu}$ . Second, we assume that the groups of mutants are monomorphic, that is, consist of a single type. This assumption implies that there is a time lag between different mutations sufficient for adjustment of the population to the original state in the case when the mutant type is weak and cannot invade.<sup>18</sup>

Next, we introduce two more definitions that are related to stability of a configuration. We call a configuration  $(\mu, b)$  *balanced* if  $\Pi_\theta(\mu, b) = \Pi_{\theta'}(\mu, b)$  for every  $\theta, \theta' \in C(\mu)$ , that is, if all constituent types have the same expected fitness in this configuration. Further, we call a configuration  $(\mu, b)$  *internally stable* if it is stable with respect to small perturbations in fractions of incumbent types. The formal definition of internal stability coincides with the definition of evolutionary (neutral) stability in everything except for the set of admissible mutant types, which is limited to  $C(\mu)$ . Obviously, neutral stability implies internal (neutral) stability, which in its turn implies balancedness of a configuration. These facts prove useful in our further analysis.

Even if neutral stability as it is defined above does not hold, a particular configuration might still be stable in a weaker sense, for example, follow some cyclical pattern. In section 1.4, we study stability of our model in a dynamic framework. In particular, we utilize the replicator dynamics model to investigate asymptotic stability and yet weaker stability notions with respect to configurations.<sup>19</sup>

### 1.3 Static framework

In this section we attempt to characterize neutrally stable configurations. We differentiate between cognitively homogeneous configurations, in which all types have the same cognitive intelligence level, and cognitively heterogeneous configurations.

**Definition 4.** A configuration  $(\mu, b)$  is *cognitively homogeneous* if  $n_\theta = n_{\theta'} \forall \theta, \theta' \in C(\mu)$ . Otherwise a configuration is *cognitively heterogeneous*.

In our setting, cognitively homogeneous configurations fall into two groups: consisting solely of naïve or of intelligent types. The former case corresponds to the models with complete unobservability of preferences, and the latter case – to the models with perfect observability. However, there is an important difference: we allow mutants to have a cognitive intelligence level different from that of incumbents.<sup>20</sup>

---

<sup>18</sup>We might consider an alternative static stability definition in which polymorphic mutations are allowed. Such a definition would be more stringent than the one we propose here, and thus every stable configuration which can be found in that case should be among stable configurations we find in this paper.

<sup>19</sup>A good overview of dynamic evolutionary models, and in particular the replicator dynamics model, can be found in Weibull (1995) or Hofbauer and Sigmund (1998).

<sup>20</sup>Hence, any stable configuration in our framework should also be stable in the framework of Dekel et al. (2007), although the opposite is not necessarily true.

The first proposition provides necessary and sufficient conditions for a cognitively homogeneous configuration with naïve players to be neutrally stable.

**Proposition 1** [COGNITIVELY HOMOGENEOUS NSC WITH NAÏVE PLAYERS]

Let  $(\mu, b)$  be a configuration with  $n_\theta = 0 \forall \theta \in C(\mu)$ . Then  $(\mu, b)$  is an NSC if and only if the following conditions hold:

(i)  $b_\theta(\theta') = D \forall \theta, \theta' \in C(\mu)$ ;

(ii) if  $f(C, D) = D$  and  $u_\theta(D, D) = u_\theta(C, D)$  for some  $\theta \in C(\mu)$ , then  $u_\theta(D, C) \geq u_\theta(C, C)$ .

The first condition of the proposition simply says that in an NSC consisting solely of naïve players everyone defects. It does not seem very surprising, as it is congruent with conclusions of the previous studies, which have shown that when preferences are not observable only Nash outcomes can be stable.

However, for each player there are two possible reasons why to choose defection in such a configuration: it can either be the consequence of a strict preference,  $u_\theta(D, D) > u_\theta(C, D)$ , or it can follow from focality of defection in the case of indifference:  $u_\theta(D, D) = u_\theta(C, D)$  and  $f(C, D) = D$ . The second condition of the proposition corresponds to the latter case and ensures that such incumbents do not immediately switch to cooperation when a cooperating mutant enters. Necessity of this condition arises from our requirement of existence of a focal post-entry equilibrium for every possible mutant.<sup>21</sup>

Let us illustrate existence of such an NSC with a simple example. Consider a population consisting of a single type  $\theta \in \Theta$ , such that  $n_\theta = 0$  and  $u_\theta = \pi$ . That is, subjective preferences of the incumbent population correspond to the actual fitness payoffs. Note that condition (i) of the proposition is satisfied, as  $(D, D)$  is a unique equilibrium in the fitness game, and since  $D$  is a strictly dominant action for all players, condition (ii) of the proposition is irrelevant.

Consider another example. Let a population consist of two naïve types:  $\theta \in \Theta$  with  $u_\theta(C, C) = u_\theta(C, D) = 1$  and  $u_\theta(D, C) = u_\theta(D, D) = 2$ , and  $\theta' \in \Theta$  with  $u_{\theta'}(C, C) = 1$  and  $u_{\theta'}(C, D) = u_{\theta'}(D, C) = u_{\theta'}(D, D) = 2$ . If the focal action (tie-breaking convention) is  $D$ , then such a population together with a behavior policy  $b_\theta(\theta') = b_{\theta'}(\theta) = D$  is also an NSC.

The next proposition characterizes cognitively homogeneous neutrally stable configurations with intelligent players. Given  $(\mu, b)$ , let us denote by  $\rho$  the share of players  $\mu(\theta \mid b_\theta(\theta') = D \forall \theta' \in \Theta \text{ s.t. } n_{\theta'} = 0 \text{ and } b_{\theta'}(\theta) = C)$ , who defect when matched with a naïve cooperating opponent, and by  $\eta$  the share  $\mu(\theta \mid b_\theta(\theta') = D \forall \theta' \in \Theta \text{ s.t. } n_{\theta'} = 0 \text{ and } b_{\theta'}(\theta) = D)$ , who defect when matched with a naïve defector.

---

<sup>21</sup>As it has been already mentioned, the requirement of existence of a focal post-entry equilibrium is relaxed in Dekel et al. (2007), and thus being a strict Nash is sufficient for stability of defection in their framework.

**Proposition 2** [COGNITIVELY HOMOGENEOUS NSC WITH INTELLIGENT PLAYERS]

Let  $(\mu, b)$  be a configuration with  $n_\theta = 1 \forall \theta \in C(\mu)$ . Then  $(\mu, b)$  is an NSC if and only if the following conditions hold:

- (i)  $b_\theta(\theta') = C \forall \theta, \theta' \in C(\mu)$ ;
- (ii)  $\rho > \frac{k}{1+x}$ ;
- (iii)  $\eta > \frac{k+y-1}{y}$ .

This result is also in line with conclusions of the related literature: when preferences are observed perfectly, only efficient outcomes are stable (in our model,  $(C, C)$  is a unique efficient outcome due to  $y - x < 2$ ). However, even though individuals in such an NSC play efficiently, they should be able to protect themselves against all potential mutant entrants.<sup>22</sup> Conditions (ii) and (iii) of the proposition guarantee that the original configuration cannot be invaded by naïve cooperating types, as well as by naïve defectors. Note, however, that  $\rho + \eta \leq 1$  implies that cognitively homogeneous NSCs exist only if  $\frac{k}{1+x} + \frac{k+y-1}{y} < 1$ , or equivalently, only if  $k < \frac{1+x}{1+x+y} (< 1)$ . If the cognitive cost paid by intelligent incumbents is too high, they will not be able to withstand competition with their naïve counterparts.

It is remarkable that cognitively homogeneous stable configurations in our framework coincide with pure stable configurations.<sup>23</sup> As Propositions 1 and 2 imply, all cognitively homogeneous NSCs are, at the same time, pure configurations, since all players in the population choose the same pure action. It is quite straightforward to show that the inverse is also true: every pure NSC requires cognitive homogeneity (otherwise it would be imbalanced).

The conclusions of Heller and Mohlin (2019) just partly coincide with our results. This discrepancy comes from differences in some basic assumptions. For example, Heller and Mohlin (2019) conclude that in a pure stable configuration everyone is of the lowest cognitive level. However, in their model increasing cognition is always possible, and once mutants are of a higher cognitive level they can use strong deception to mislead their opponents and invade the population. In our framework, a configuration with intelligent players may also be stable, as there is no one who could surpass the incumbents in their cognitive intelligence.

Furthermore, according to Heller and Mohlin (2019), a necessary condition for stability of a pure configuration is that the outcome should be efficient. In our model, however, the outcome in a pure NSC with naïve players is not efficient. This difference can be explained

---

<sup>22</sup>Note that even though naïve types are not present in the configuration, they can enter as mutants. A stable configuration should be proof against such mutants too. In the setting of Dekel et al. (2007), this was not the case, that is why they do not need additional conditions to be satisfied for stability.

<sup>23</sup>Following Heller and Mohlin (2019), a configuration  $(\mu, b)$  is *pure* if there exists an action  $a^* \in A$  such that  $b_\theta(\theta') = a^*$  for every  $\theta, \theta' \in C(\mu)$ .

by the fact that Heller and Mohlin (2019) allow for polymorphic groups of mutants, while we assume only one mutant type entering at a time. It follows that, in our framework, efficient play is not crucial for stability of a homogeneous configuration with naïve types, since no single mutant type can outdo the incumbents on its own.

Finally, we formulate necessary conditions for neutral stability of a cognitively heterogeneous configuration. To simplify the exposition, let us introduce some additional notation: given  $(\mu, b)$ , we denote by  $\tau$  the share of players  $\mu(\theta \mid n_\theta = 0$  and  $b_\theta(\theta') = C \forall \theta' \in C(\mu)$ ), who are naïve and cooperate in  $(\mu, b)$ , by  $\omega^c$  the share of players  $\mu(\theta \mid n_\theta = 1$  and  $b_\theta(\theta') = b_{\theta'}(\theta) \forall \theta' \in \Theta$  s.t.  $n_{\theta'} = 0$ ), who are intelligent and always match the action of their naïve opponent, and by  $\omega^a$  the share of players  $\mu(\theta \mid n_\theta = 1$  and  $b_\theta(\theta') \neq b_{\theta'}(\theta) \forall \theta' \in \Theta$  s.t.  $n_{\theta'} = 0$ ), who are intelligent and always mismatch the action of their naïve opponent.<sup>24</sup>

**Proposition 3** [NECESSARY CONDITIONS FOR COGNITIVELY HETEROGENEOUS NSC]

Let  $(\mu, b)$  be an NSC such that  $\exists \theta, \theta' \in C(\mu)$  with  $n_\theta = 0$  and  $n_{\theta'} = 1$ . Then the following conditions must hold:

- (i)  $\omega^a < \frac{1}{1+x+y}$ ;
- (ii)  $\omega^c > \frac{1+x}{y} \cdot \omega^a + \frac{y-1}{y}$ ;
- (iii)  $\tau + \omega^c + \omega^a = 1$ .

It appears interesting that cognitively heterogeneous stable configurations have a quite restricted set of constituent types. In essence, all naïve individuals cooperate, and intelligent players prefer either to match or to mismatch the actions of their naïve opponents. Naïve types play efficiently among themselves and with a fraction of intelligent types, while earning the lowest fitness payoff when matched with the remaining fraction of intelligent types. Intelligent players pay a cognitive cost for the possibility to tailor their actions to their opponents, which might help them to play efficiently in some matches while avoiding the lowest fitness payoff in other matches. Since the two classes of intelligent types earn different fitness payoffs with the naïve fraction of the population, they must also play differently among themselves in order to preserve balancedness of the configuration.

Note that the above conditions are necessary but not sufficient for neutral stability of a configuration. On the other hand, many other preferences that do not appear in stable configurations can still exist temporarily during the evolutionary process. Thus, in the next section we consider the evolution of preferences from the dynamic perspective.

---

<sup>24</sup>The superscripts in  $\omega^c$  and  $\omega^a$  stand for "coordination" and "anti-coordination" of own action with the action of a naïve opponent.

## 1.4 Dynamic framework

In this section, we narrow down our attention to the particular set of admissible social preferences, which are both easily interpretable and, according to our previous analysis, seem likely to appear in a configuration demonstrating stability properties.

Suppose, there are four possible social preferences (subjective preferences over outcomes in the prisoner’s dilemma game):

$u^D$  – ”defectors”, who strictly prefer to defect,

$u^C$  – ”cooperators”, who strictly prefer to cooperate,

$u^c$  – ”matchers”, or ”conformists”, who strictly prefer to use the same action as their opponent, and

$u^a$  – ”mismatchers”, or ”anti-conformists”, who strictly prefer to use the opposite action.<sup>25</sup>

Having added the cognitive component, we get the following type space consisting of eight possible types:  $\Theta = \{\theta_0^D, \theta_0^C, \theta_0^c, \theta_0^a, \theta_1^D, \theta_1^C, \theta_1^c, \theta_1^a\}$ .<sup>26</sup> Note that types  $\theta_1^D$  and  $\theta_1^C$  will never appear in a stable configuration, since in any type game they are strictly dominated by the types  $\theta_0^D$  and  $\theta_0^C$  respectively (this is true for any type pre-committed to a specific strategy). Thus, we are left to analyze stability of configurations with the following type space:  $\Theta = \{\theta_0^D, \theta_0^C, \theta_0^c, \theta_0^a, \theta_1^c, \theta_1^a\}$ .

We use computer simulations to investigate dynamic stability of different cognitively heterogeneous configurations. To describe the dynamic process, we apply a classical dynamic model of evolutionary selection in continuous time – the replicator dynamics. In this model mutations are not introduced explicitly; instead, the selection mechanism determines how the population state (the type distribution) evolves over time. Varying the set of admissible types and their behaviors towards each other (in the case of multiple equilibria), we check whether the population state stabilizes over time for different values of the cognitive cost.

First of all, we check stability properties of the configuration that satisfies necessary conditions for neutral stability (Proposition 3). It includes the following types:  $\theta_0^C$ ,  $\theta_1^c$  and  $\theta_1^a$ . As Figure 1.1 shows, the population state stabilizes over time at different levels, depending on the value of the cognitive cost  $k$  (or fluctuates around these levels). The fraction of each of the constituent types is strictly positive and significant, which supports the conclusions of section 1.3.

---

<sup>25</sup>Note that defectors and cooperators here are pre-committed to play a specific action unconditionally.

<sup>26</sup>For simplicity, in this section we adopt a slightly modified notation, where the subscript of a type corresponds to its cognitive component and the superscript – to its preference component.



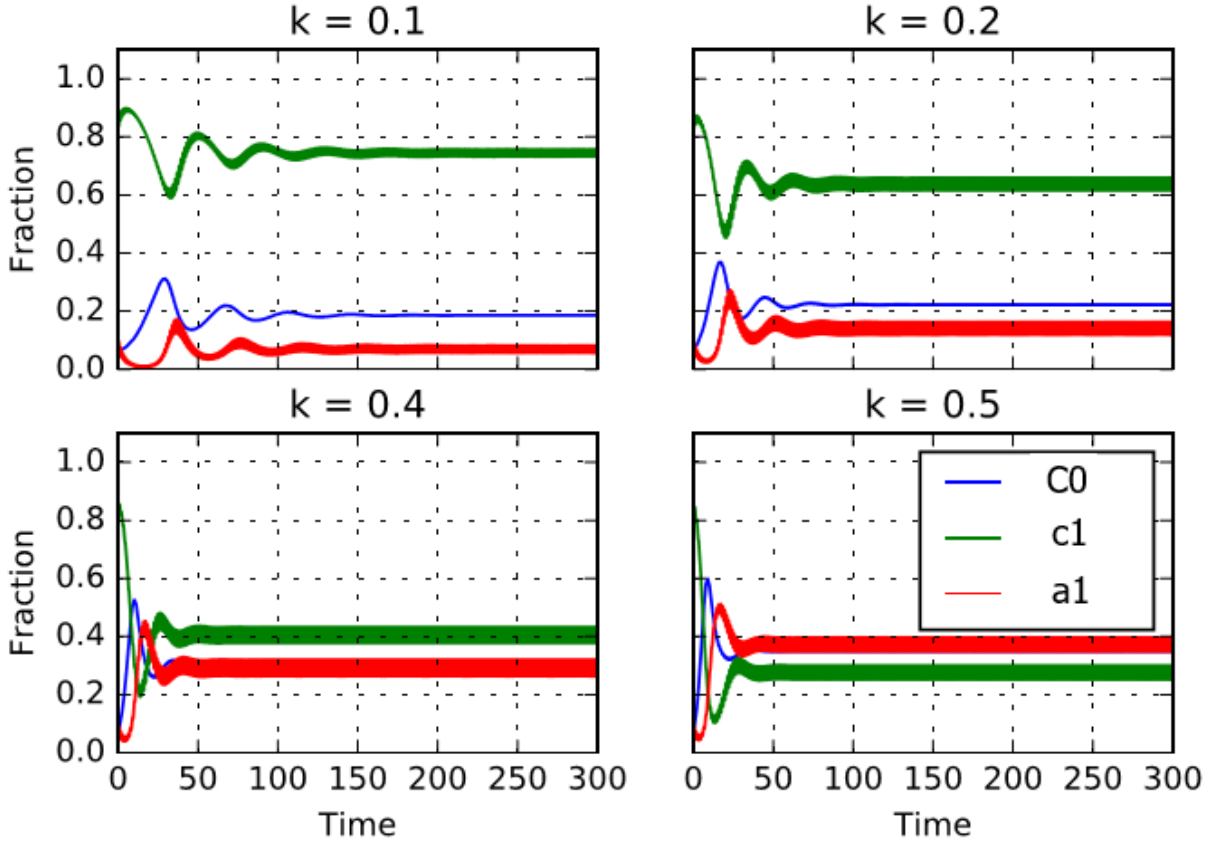


Figure 1.1: Dynamics of the population state for the configuration consisting of naïve cooperators ( $C0$ ), intelligent matchers ( $c1$ ) and intelligent mismatchers ( $a1$ )

If we add type  $\theta_0^D$  to the configuration, the picture changes dramatically. For a lower cognitive cost (Figure 1.2, left), population state never stabilizes; however, it follows a regular cyclical pattern with positive fractions of all types attained periodically. For a higher cognitive cost (Figure 1.2, right), type  $\theta_0^D$  outperforms all the other types, driving their shares to zero: intelligent competitors lose because of the high cognitive cost and, in their absence,  $\theta_0^C$  becomes weaker than  $\theta_0^D$  due to strict dominance of defection.

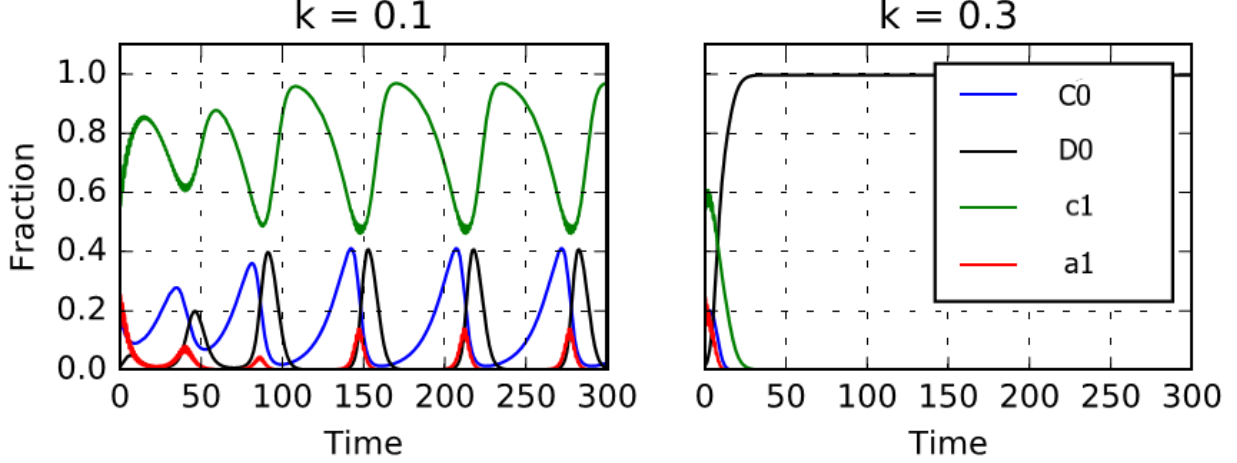


Figure 1.2: Dynamics of the population state for the configuration consisting of naïve cooperators (C0), naïve defectors (D0), intelligent matchers (c1) and intelligent mismatchers (a1)

## 1.5 Extension: Public preference and deception

Below we outline one interesting extension of our model. We introduce an additional component to a type of an individual – a *public preference* – that can be observed by everyone regardless of their cognitive intelligence level. However, we assume that only intelligent types can have a public preference different from their *private (true) preference*, or to put it differently, only intelligent types can deceive others about their true preferences. Naïve types are assumed not capable of preference mimicry, thus their public and private preferences always coincide.

Hence, in this extended model cognitive ability has two dimensions – the observational and the imitational one. Note that not only intelligent types get an additional advantage in this setting; naïve types also improve their standing with respect to the information they have, as now they can discriminate their opponents on the basis of their public preferences.

Formally, a *type* of a player is now a triple, consisting of a subjective preference, a public preference (public image) and a cognitive intelligence level:  $\theta = (u, \hat{u}, n) \in \Theta = U^2 \times \{0, 1\}$ , and  $\hat{u}_\theta = u_\theta$  whenever  $n_\theta = 0$ . We define a *configuration* as a pair  $(\mu, b)$ , where  $\mu \in \Delta(\Theta)$  is a type distribution with finite support  $C(\mu)$  and  $b : C(\mu) \times C(\mu) \rightarrow A$  is a behavior policy such that for every  $\theta, \theta' \in C(\mu)$ :

$$b(\theta, \theta') = b_\theta(\theta') = \begin{cases} f(BR_\theta(b_{\theta'}(\theta))) & \text{if } n_\theta = 1, \\ f(BR_\theta(\sum_{\{\theta_i | \hat{u}_i = \hat{u}'\}} \frac{\mu(\theta_i)}{\sum_{\{\theta_j | \hat{u}_j = \hat{u}'\}} \mu(\theta_j)} b_{\theta_i}(\theta))) & \text{if } n_\theta = 0. \end{cases} \quad (1.2)$$

Since a naïve player can observe the public preference of her opponent, she best-responds to the average play of all those who have the same public preference. An intelligent type, as

before, observes perfectly her opponent’s type and best-plies to her behavior.

An essential difference of this setting from the one we consider in our paper is that it allows for a qualitatively new phenomenon – deception. Intelligent players might try to use deception in order to manipulate behavior of their opponents in a way favorable for the deceivers. Naïve players are aware of the possibility of deception, even though they cannot observe true preferences of their opponents before the play. Intuitively, the benefit from deception should decline when a higher fraction of the population uses public images different from their true preferences. Is there an equilibrium level of deception corresponding to the static stability solution? If not, which dynamics does it follow? And how does it correspond to different distributions of social preferences in the population? It would be interesting to try to answer these questions within our extended framework.

It is worth noting, however, that in this new framework observational and imitational abilities of individuals are pooled. Yet another natural model extension would be to separate these cognitive abilities. We could let a *type* of a player include four independent components – two preference components, as before, and two cognitive ones:  $\theta = (u, \hat{u}, n, m) \in \Theta = U^2 \times \{0, 1\}^2$ . The first two components would correspond to the true subjective preference and the public preference of an individual, the third component would determine whether an individual is able to observe her opponent’s true preference, and the fourth component would determine her ability to fake own preferences. The behavior policy in this setting would be still given by (1.2), but we will have to assume that  $\hat{u}_\theta = u_\theta$  whenever  $m_\theta = 0$ .

Generally, we should also assume two different costs for two separate cognitive abilities, since now these abilities can develop independently. Intuitively, the ability to see through true preferences of others should be correlated with the ability to disguise own preferences. To test this correlation within the proposed framework seems to be an interesting endeavor.

## 1.6 Conclusions

We propose a model in which social preferences coevolve together with the level of their observability. Since observability is endogenous in this model, our analysis can cover the whole range of partial observability cases. Players are characterized by their preferences over the outcomes in a two-player game and a level of their cognitive intelligence, which determines their ability to observe the preferences of their opponents. The cost of such information is strictly positive, which rules out the possibility of a ”secret handshake”.

We provide complete characterization of the set of neutrally stable configurations of players of the same cognitive intelligence level, and find that our results are consistent with those of other related studies. Furthermore, we derive necessary conditions for neutral stability of cognitively heterogeneous configurations. These conditions suggest that the class of social

preferences that can appear in stable configurations is quite restricted.

Apart from theoretical static solution results, the dynamic solution framework has been analyzed. Dynamic simulations have revealed that some cognitively heterogeneous configurations have properties very close to stable. Some other configurations follow a particular cyclical pattern, which can be considered as stability in a weaker sense.

## Appendix to Chapter 1

### Proof of Proposition 1

Let  $(\mu, b)$  be such that  $n_\theta = 0 \forall \theta \in C(\mu)$ . Then the definition of a behavior policy implies that for any  $\theta, \theta', \theta'' \in C(\mu)$ ,  $b_\theta(\theta') = b_\theta(\theta'') = f(BR_\theta(b \cdot (\theta); \mu))$ . With a slight abuse of notation, for each  $\theta \in C(\mu)$  let us denote this action by  $b_\theta(\mu)$ .

Necessity. Let  $(\mu, b)$  be an NSC and suppose condition (i) does not hold, i.e.  $\exists \theta \in C(\mu)$  such that  $b_\theta(\mu) = C$ . If  $\mu(\theta \in C(\mu) \mid b_\theta(\mu) = C) < 1$ , then  $\exists \theta' \in C(\mu)$  s.t.  $b_{\theta'}(\mu) = D$  and hence  $\Pi_{\theta'}(\mu, b) > \Pi_\theta(\mu, b)$ . In this case  $(\mu, b)$  is not balanced and thus cannot be an NSC. If  $\mu(\theta \in C(\mu) \mid b_\theta(\mu) = C) = 1$ , then a mutant type  $\theta' \in \Theta$  with  $n_{\theta'} = 0$  and  $b_{\theta'}(\tilde{\mu}) = D$  can invade:  $\tilde{b}_{\theta'}(\tilde{\mu}) = b_\theta(\mu) = C \forall \theta \in C(\mu)$  implies that  $\Pi_{\theta'}(\tilde{\mu}, \tilde{b}) > \Pi_\mu(\tilde{\mu}, \tilde{b})$ . Again,  $(\mu, b)$  cannot be an NSC. Thus, condition (i) must hold.

Now let condition (ii) fail, i.e.  $f(C, D) = D$  and  $\exists \theta \in C(\mu)$  s.t.  $u_\theta(D, D) = u_\theta(C, D)$  and  $u_\theta(D, C) < u_\theta(C, C)$ . Consider a mutant type  $\theta' \in \Theta$  with  $n_{\theta'} = 0$ ,  $u_{\theta'}(D, D) < u_{\theta'}(C, D)$  and  $u_{\theta'}(D, C) < u_{\theta'}(C, C)$ . Since  $C$  is a strictly dominant action for  $\theta'$ , it must be that  $\tilde{b}_{\theta'}(\tilde{\mu}) = C$  in every post-entry configuration  $(\tilde{\mu}, \tilde{b})$ . And since  $\tilde{\mu}(\theta'' \mid \tilde{b}_{\theta''}(\tilde{\mu}) = C) > 0$ , it must be that  $\tilde{b}_\theta(\tilde{\mu}) = C$  for the incumbent type  $\theta$ , implying that no focal post-entry configuration exists and, consequently, the original configuration  $(\mu, b)$  is not an NSC. Thus, condition (ii) must also hold.

Sufficiency. Let  $(\mu, b)$  be such that both conditions hold. Condition (i) implies that  $\Pi_\theta(\mu, b) = \Pi_{\theta'}(\mu, b) \forall \theta, \theta' \in C(\mu)$ , and thus  $(\mu, b)$  is balanced and internally stable. We are left to check that no mutant type  $\theta^* \notin C(\mu)$  can either invade the population or destabilize it by changing the incumbent behavior.

Let us first prove the existence of a focal post-entry configuration for any  $\theta^* \notin C(\mu)$ . Consider an incumbent of type  $\theta \in C(\mu)$ . If  $\tilde{b}_{\theta^*}(\theta) = D$ , then  $\tilde{b}_\theta(\tilde{\mu}) = b_\theta(\mu) = D$ , that is, the behavior of the incumbent is unchanged. Suppose now that  $\tilde{b}_{\theta^*}(\theta) = C$ . Then  $u_\theta(D, (\varepsilon, 1 - \varepsilon)) - u_\theta(C, (\varepsilon, 1 - \varepsilon)) = \varepsilon (u_\theta(D, C) - u_\theta(C, C)) + (1 - \varepsilon)(u_\theta(D, D) - u_\theta(C, D))$ . This means that if  $u_\theta(D, D) - u_\theta(C, D) > \frac{\varepsilon}{1 - \varepsilon} (u_\theta(C, C) - u_\theta(D, C))$ , then  $u_\theta(D, (\varepsilon, 1 - \varepsilon)) > u_\theta(C, (\varepsilon, 1 - \varepsilon))$ , and thus  $\tilde{b}_\theta(\tilde{\mu}) = D$ . There are two possibilities. If  $\theta$  is such that  $u_\theta(D, D) > u_\theta(C, D)$ , we can always find sufficiently small  $\bar{\varepsilon} > 0$  such that  $u_\theta(D, D) - u_\theta(C, D) > \frac{\varepsilon}{1 - \varepsilon} (u_\theta(C, C) - u_\theta(D, C)) \forall \varepsilon < \bar{\varepsilon}$ . If, however,  $\theta$  is

such that  $u_\theta(D, D) = u_\theta(C, D)$  and  $f(C, D) = D$ , then condition (ii) guarantees that  $u_\theta(D, C) \geq u_\theta(C, C)$ . Then  $u_\theta(D, D) - u_\theta(C, D) = 0 \geq \frac{\varepsilon}{1-\varepsilon} (u_\theta(C, C) - u_\theta(D, C)) \forall \varepsilon > 0$ , and hence  $u_\theta(D, (\varepsilon, 1 - \varepsilon)) \geq u_\theta(C, (\varepsilon, 1 - \varepsilon))$ , which together with  $f(C, D) = D$  implies  $\tilde{b}_\theta(\tilde{\mu}) = D$ . Thus, in any case a focal post-entry configuration exists.

Now let  $(\tilde{\mu}, \tilde{b})$  be a focal post-entry configuration, i.e.  $\tilde{b}_\theta(\tilde{\mu}) = b_\theta(\mu) = D \forall \theta \in C(\mu)$ . If  $n_{\theta^*} = 0$ , then  $\Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) \leq \Pi_\mu(\tilde{\mu}, \tilde{b})$ , due to strict dominance of  $D$  in the fitness game. If  $n_{\theta^*} = 1$ , then its fitness payoff is maximal if  $\tilde{b}_{\theta^*}(\theta^*) = C$  and  $\tilde{b}_{\theta^*}(\theta) = D \forall \theta \in C(\mu)$ . In this case  $\Pi_\mu(\tilde{\mu}, \tilde{b}) = (1-\varepsilon)\pi(D, D) + \varepsilon\pi(D, D) = 0$  and  $\Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) = (1-\varepsilon)\pi(D, D) + \varepsilon\pi(C, C) - k = \varepsilon - k$ . Obviously,  $\forall k > 0 \exists \bar{\varepsilon}(k) = k$  s.t.  $\forall \varepsilon < \bar{\varepsilon} : \Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) < \Pi_\mu(\tilde{\mu}, \tilde{b})$ . Thus, no mutant, naïve or intelligent, can invade, and  $(\mu, b)$  is an NSC.  $\square$

## Proof of Proposition 2

Let  $(\mu, b)$  be such that  $n_\theta = 1 \forall \theta \in C(\mu)$ .

Necessity. Suppose that  $(\mu, b)$  is an NSC and let us prove the necessity of all three conditions in turn.

First, suppose that condition (i) does not hold, i.e.  $\exists \theta, \theta' \in C(\mu)$  such that  $b_\theta(\theta') = D$ , and let us show that in this case  $(\mu, b)$  is either imbalanced or can be invaded by a naïve mutant. If  $\exists \theta'' \in C(\mu)$  s.t.  $b_{\theta''}(\theta') = C \forall \theta' \in C(\mu)$ , then  $\Pi_{\theta''}(\mu, b) < \Pi_\theta(\mu, b)$ , implying that  $(\mu, b)$  is imbalanced and thus not an NSC. Suppose now that  $b_\theta(\theta') = D \forall \theta, \theta' \in C(\mu)$ . Consider a mutant type  $\theta^* \in \Theta$  s.t.  $n_{\theta^*} = 0$  and  $b_{\theta^*}(\tilde{\mu}) = D$ .<sup>27</sup> This mutant can invade the population, as  $\Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) - \Pi_\mu(\tilde{\mu}, \tilde{b}) = k > 0$ , which again contradict neutral stability of  $(\mu, b)$ . Thus, it must be that  $b_\theta(\theta') = C \forall \theta, \theta' \in C(\mu)$ .

Now suppose that condition (ii) fails, i.e.  $\rho \leq \frac{k}{1+x}$ . Consider a mutant type  $\theta^* \in \Theta$  s.t.  $n_{\theta^*} = 0$  and  $b_{\theta^*}(\tilde{\mu}) = C$ . The expected fitness payoff of the mutant type is  $\Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)(\rho \pi(C, D) + (1 - \rho)\pi(C, C)) + \varepsilon \pi(C, C) = 1 - (1 - \varepsilon)\rho(1 + x)$ , and the expected fitness payoff of the incumbent population is  $\Pi_\mu(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)\pi(C, C) + \varepsilon (\rho \pi(D, C) + (1 - \rho)\pi(C, C)) - k = 1 + \varepsilon\rho(y - 1) - k$ . It follows that  $\Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) > \Pi_\mu(\tilde{\mu}, \tilde{b})$  if and only if  $\rho < \frac{k}{(1-\varepsilon)(1+x)+\varepsilon(y-1)}$ . As  $1 + x > y - 1$  implies  $\frac{k}{1+x} < \frac{k}{(1-\varepsilon)(1+x)+\varepsilon(y-1)} \forall \varepsilon \in (0; 1]$ , it follows that  $\rho \leq \frac{k}{1+x}$  implies  $\rho < \frac{k}{(1-\varepsilon)(1+x)+\varepsilon(y-1)} \forall \varepsilon > 0$ , and thus there is no invasion barrier for the mutant type  $\theta^*$ . Therefore, if  $(\mu, b)$  is an NSC,  $\rho > \frac{k}{1+x}$  must hold.

Finally, suppose that condition (iii) fails, i.e.  $\eta \leq \frac{k+y-1}{y}$ . A similar reasoning applies in this case. Consider a mutant type  $\theta^* \in \Theta$  s.t.  $n_{\theta^*} = 0$  and  $b_{\theta^*}(\tilde{\mu}) = D$ . The expected fitness payoff of the mutant type is  $\Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)(\eta \pi(D, D) + (1 - \eta)\pi(D, C)) + \varepsilon \pi(D, D) = (1 - \varepsilon)(1 - \eta) y$ , and the expected fitness payoff of the incumbent population is  $\Pi_\mu(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)\pi(C, C) + \varepsilon(\eta \pi(D, D) + (1 - \eta)\pi(C, D)) - k = (1 - \varepsilon) - \varepsilon(1 - \eta) x - k$ . It follows that

<sup>27</sup>As it has been shown in the proof of Proposition 1, for any configuration  $(\mu, b)$  and any type  $\theta \in \Theta$  with  $n_\theta = 0$  it holds that  $b_\theta(\theta') = b_\theta(\theta'') \forall \theta', \theta'' \in C(\mu)$ . Therefore, we can use the simplified notation  $b_\theta(\mu)$  for the action of  $\theta$  in  $(\mu, b)$ .

$\Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) > \Pi_{\mu}(\tilde{\mu}, \tilde{b})$  if and only if  $\eta < \frac{k+y-1+\varepsilon(1+x-y)}{y+\varepsilon(x-y)}$ . However,  $\eta \leq \frac{k+y-1}{y}$  implies that  $\eta < \frac{k+y-1+\varepsilon(1+x-y)}{y+\varepsilon(x-y)} \forall \varepsilon > 0$ , and thus there is no invasion barrier for the mutant type  $\theta^*$ . (To prove this last implication, observe that when  $k(x-y) < x$ , then  $\frac{k+y-1}{y} < \frac{k+y-1+\varepsilon(1+x-y)}{y+\varepsilon(x-y)} \forall \varepsilon > 0$ . And when  $k(x-y) \geq x$ , then  $\frac{k+y-1}{y} \geq \frac{k+y-1+\varepsilon(1+x-y)}{y+\varepsilon(x-y)} > 1 \forall \varepsilon > 0$ , where the last inequality follows from  $k \geq \frac{x}{x-y} > 1$ .) Therefore, if  $(\mu, b)$  is an NSC,  $\eta > \frac{k+y-1}{y}$  must hold.

Thus, all three conditions are necessary for neutral stability of  $(\mu, b)$ .

Sufficiency. Suppose that  $(\mu, b)$  satisfies conditions (i), (ii) and (iii) and let us prove that it is an NSC. Since  $n_{\theta} = n_{\theta'}$  and, due to condition (i),  $b_{\theta}(\theta'') = b_{\theta'}(\theta'') \forall \theta, \theta', \theta'' \in C(\mu)$ , the configuration is balanced and internally stable. We are left to check that no mutant type  $\theta^* \notin C(\mu)$  can invade.

Consider a mutant type  $\theta^* \in \Theta$  s.t.  $n_{\theta^*} = 0$  and  $b_{\theta^*}(\tilde{\mu}) = C$ . As it follows from the necessity part of the proof,  $(\mu, b)$  is proof against such a mutant if and only if  $\exists \bar{\varepsilon} > 0$  s.t.  $\rho \geq \frac{k}{(1-\varepsilon)(1+x)+\varepsilon(y-1)} \forall \varepsilon < \bar{\varepsilon}$ . However, since  $\rho > \frac{k}{1+x}$  (condition (ii)), such an invasion barrier indeed exists:  $\bar{\varepsilon} = \frac{1+x-k/\rho}{2-y+x} > 0$ . As the value of  $\frac{k}{(1-\varepsilon)(1+x)+\varepsilon(y-1)}$  strictly increases in  $\varepsilon$ , it follows that  $\rho = \frac{k}{(1-\bar{\varepsilon})(1+x)+\bar{\varepsilon}(y-1)} > \frac{k}{(1-\varepsilon)(1+x)+\varepsilon(y-1)} \forall \varepsilon < \bar{\varepsilon}$ . Thus,  $(\mu, b)$  is proof against such a mutation.

Consider now a mutant type  $\theta^* \in \Theta$  s.t.  $n_{\theta^*} = 0$  and  $b_{\theta^*}(\tilde{\mu}) = D$ . It follows from the necessity part of the proof that  $(\mu, b)$  is proof against such a mutant if and only if  $\exists \bar{\varepsilon} > 0$  s.t.  $\eta \geq \frac{k+y-1+\varepsilon(1+x-y)}{y+\varepsilon(x-y)} \forall \varepsilon < \bar{\varepsilon}$ . Two cases have to be considered separately. First, if  $k(x-y) < x$ , then together with  $\eta > \frac{k+y-1}{y}$  (condition (iii)) it implies that such a (positive) invasion barrier indeed exists:  $\bar{\varepsilon} = \frac{(1-\eta)y+k-1}{(1-\eta)(y-x)-1} > 0$ . As the value of  $\frac{k+y-1+\varepsilon(1+x-y)}{y+\varepsilon(x-y)}$  strictly increases in  $\varepsilon$  in this case, it follows that  $\eta = \frac{k+y-1+\bar{\varepsilon}(1+x-y)}{y+\bar{\varepsilon}(x-y)} > \frac{k+y-1+\varepsilon(1+x-y)}{y+\varepsilon(x-y)} \forall \varepsilon < \bar{\varepsilon}$ . Second, if  $k(x-y) \geq x$ , then the value of  $\frac{k+y-1+\varepsilon(1+x-y)}{y+\varepsilon(x-y)}$  is non-increasing in  $\varepsilon$ , and hence  $\eta > \frac{k+y-1}{y} \geq \frac{k+y-1+\varepsilon(1+x-y)}{y+\varepsilon(x-y)} \forall \varepsilon > 0$ . In any case,  $(\mu, b)$  is proof against such a deviation.

Finally, consider a mutant type  $\theta^* \notin C(\mu)$  with  $n_{\theta^*} = 1$ . Note that this is the case with complete observability of types and  $k_{\theta^*} = k_{\theta} \forall \theta \in C(\mu)$ . Since the incumbents already play the only efficient pure strategy equilibrium, in a focal post-entry configuration  $\Pi_{\mu}(\tilde{\mu}, \tilde{b}) \geq \Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) \forall u_{\theta^*} \in U$ . Since  $(\mu, b)$  is proof against all possible mutations by both naïve and intelligent players, it is an NSC.  $\square$

### Proof of Proposition 3

Let us first prove several auxiliary lemmas.

**Lemma 1.** *Let  $(\mu, b)$  be a cognitively heterogeneous NSC. If  $n_{\theta} = 0$ , then  $b_{\theta}(\theta') = C \forall \theta' \in C(\mu)$ .*

*Proof.* Let  $(\mu, b)$  be a cognitively heterogeneous NSC. It implies that  $(\mu, b)$  is internally neutrally stable. Assume that  $\exists \theta \in C(\mu)$  with  $n_{\theta} = 0$  and  $b_{\theta}(\mu) = D$  (see footnote 27) and let the type  $\theta$  increase its share in  $\mu$ .

Since  $(\mu, b)$  is internally neutrally stable, there must exist a focal post-entry configuration:  $\tilde{\mu} = (1 - \varepsilon)\mu + \varepsilon\theta$  and  $\tilde{b}(\theta', \theta'') = b(\theta', \theta'') \forall \theta', \theta'' \in C(\tilde{\mu})$  (note that  $C(\tilde{\mu}) = C(\mu)$ ). Since the original configuration  $(\mu, b)$  is an NSC, it is balanced:  $\Pi_\theta(\mu, b) = \Pi_{\theta'}(\mu, b) \forall \theta' \in C(\mu)$ . The fitness payoff of  $\theta$  in the post-entry configuration is  $\Pi_\theta(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)\Pi_\theta(\mu, b) + \varepsilon \cdot 0 = (1 - \varepsilon)\Pi_\theta(\mu, b)$ . At the same time, the fitness payoff of any other naïve type  $\theta' \in C(\mu)$  with  $b_{\theta'}(\mu) = D$  is the same, i.e.  $\Pi_{\theta'}(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)\Pi_\theta(\mu, b)$ , and if  $b_{\theta'}(\mu) = C$ , it is even lower:  $\Pi_{\theta'}(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)\Pi_\theta(\mu, b) + \varepsilon \cdot (-x)$ . The fitness payoff of any intelligent type  $\theta' \in C(\mu)$  in the post-entry configuration is always lower than that of  $\theta$ : if  $b_{\theta'}(\theta) = D$ , then  $\Pi_{\theta'}(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)(\Pi_\theta(\mu, b) + k) + \varepsilon \cdot 0 - k = (1 - \varepsilon)\Pi_\theta(\mu, b) - \varepsilon k$ , and if  $b_{\theta'}(\theta) = C$ , then  $\Pi_{\theta'}(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)(\Pi_\theta(\mu, b) + k) + \varepsilon \cdot (-x) - k = (1 - \varepsilon)\Pi_\theta(\mu, b) - \varepsilon x - \varepsilon k$ . Thus, in the post-entry configuration  $\Pi_\theta(\tilde{\mu}, \tilde{b}) \geq \Pi_{\theta'}(\tilde{\mu}, \tilde{b}) \forall \theta' \in C(\tilde{\mu})$ , and moreover,  $\Pi_\theta(\tilde{\mu}, \tilde{b}) > \Pi_{\theta'}(\tilde{\mu}, \tilde{b}) \forall \theta' \in C(\tilde{\mu})$  with  $n_{\theta'} = 1$ .

Since  $(\mu, b)$  is cognitively heterogeneous,  $\mu(\theta' \mid n_{\theta'} = 1) > 0$ , and consequently,  $\Pi_\theta(\tilde{\mu}, \tilde{b}) > \Pi_{\mu'}(\tilde{\mu}, \tilde{b})$ . This implies that  $(\mu, b)$  is not internally stable. We got a contradiction, thus proving that  $n_\theta = 0$  must imply  $b_\theta(\mu) = C$ .  $\square$

**Lemma 2.** *Let  $(\mu, b)$  be a cognitively heterogeneous NSC and  $\theta \in C(\mu)$  with  $n_\theta = 1$ . Then either  $b_\theta(\theta') = b_{\theta'}(\theta) \forall \theta' \in \Theta$  s.t.  $n_{\theta'} = 0$ , or  $b_\theta(\theta') \neq b_{\theta'}(\theta) \forall \theta' \in \Theta$  s.t.  $n_{\theta'} = 0$ .*

*Proof.* Let  $(\mu, b)$  be a cognitively heterogeneous NSC. To simplify the exposition, denote by  $\Theta_0^C$  the set of all such  $\theta \in C(\mu)$  that  $n_\theta = 0$  and  $b_\theta(\theta') = C \forall \theta' \in C(\mu)$ . Similarly, denote by  $\Theta_0^D$  the set of such  $\theta \in C(\mu)$  that  $n_\theta = 0$  and  $b_\theta(\theta') = D \forall \theta' \in C(\mu)$ . Then every intelligent type in  $(\mu, b)$  falls into one of the following classes (conditional on its behavior towards naïve cooperators and naïve defectors):

$$\begin{aligned} \Theta_1^{CD} &= \{\theta \in C(\mu) \mid n_\theta = 1 \text{ and } b_\theta(\theta') = C \forall \theta' \in \Theta_0^C \text{ and } b_\theta(\theta') = D \forall \theta' \in \Theta_0^D\}, \\ \Theta_1^{DC} &= \{\theta \in C(\mu) \mid n_\theta = 1 \text{ and } b_\theta(\theta') = D \forall \theta' \in \Theta_0^C \text{ and } b_\theta(\theta') = C \forall \theta' \in \Theta_0^D\}, \\ \Theta_1^{CC} &= \{\theta \in C(\mu) \mid n_\theta = 1 \text{ and } b_\theta(\theta') = C \forall \theta' \in \Theta_0^C \cup \Theta_0^D\}, \\ \Theta_1^{DD} &= \{\theta \in C(\mu) \mid n_\theta = 1 \text{ and } b_\theta(\theta') = D \forall \theta' \in \Theta_0^C \cup \Theta_0^D\}. \end{aligned}$$

Note that  $\Pi_\theta(\mu, b) < \Pi_{\theta'}(\mu, b) \forall \theta \in \Theta_1^{CC} \forall \theta' \in \Theta_0^C$  with  $u_\theta = u_{\theta'}$ . Similarly,  $\Pi_\theta(\mu, b) < \Pi_{\theta'}(\mu, b) \forall \theta \in \Theta_1^{DD} \forall \theta' \in \Theta_0^D$  with  $u_\theta = u_{\theta'}$ . That is, these intelligent types lose in fitness terms to the corresponding naïve types. Thus, in an NSC every  $\theta \in C(\mu)$  with  $n_\theta = 1$  belongs either to  $\Theta_1^{CD}$  or to  $\Theta_1^{DC}$ . In other words, either  $b_\theta(\theta') = b_{\theta'}(\theta) \forall \theta' \in \Theta$  s.t.  $n_{\theta'} = 0$ , or  $b_\theta(\theta') \neq b_{\theta'}(\theta) \forall \theta' \in \Theta$  s.t.  $n_{\theta'} = 0$ .  $\square$

Now let us prove the proposition. Suppose  $(\mu, b)$  be an NSC, such that  $\exists \theta, \theta' \in C(\mu)$  with  $n_\theta = 0$  and  $n_{\theta'} = 1$ . Recall that  $\tau := \mu(\theta \mid n_\theta = 0 \text{ and } b_\theta(\theta') = C \forall \theta' \in C(\mu))$ ,  $\omega^c := \mu(\theta \mid n_\theta = 1 \text{ and } b_\theta(\theta') = b_{\theta'}(\theta) \forall \theta' \in \Theta \text{ s.t. } n_{\theta'} = 0)$ , and  $\omega^a := \mu(\theta \mid n_\theta = 1 \text{ and } b_\theta(\theta') \neq b_{\theta'}(\theta) \forall \theta' \in \Theta \text{ s.t. } n_{\theta'} = 0)$ . Note that condition (iii) of the proposition directly follows from

Lemmas 1 and 2:  $\tau + \omega^c + \omega^a = 1$  in every cognitively heterogeneous NSC. Let us prove the remaining two conditions.

Since  $(\mu, b)$  is balanced,  $\Pi_\mu(\mu, b) = \Pi_\theta(\mu, b) \forall \theta \in C(\mu)$ , and in particular,  $\forall \theta \in C(\mu)$  with  $n_\theta = 0$ . We can use condition (iii) to calculate this fitness payoff:  $\Pi_\mu(\mu, b) = (\tau + \omega^c) \cdot 1 + \omega^a \cdot (-x) = 1 - \omega^a(1 + x)$ .

Consider a mutant type  $\theta^*$  s.t.  $n_{\theta^*} = 1$  and  $b_{\theta^*}(\tilde{\mu}) = D$  in a focal post-entry configuration  $(\tilde{\mu}, \tilde{b})$ . The expected fitness payoff of the mutant type is  $\Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) = (1 - \varepsilon)((\tau + \omega^a)y + \omega^c \cdot 0) + \varepsilon \cdot 0 = (1 - \varepsilon)(1 - \omega^c)y$ , and the expected fitness payoff of the incumbent population is  $\Pi_\mu(\tilde{\mu}, \tilde{b}) = \tau[(1 - \varepsilon)\Pi_\mu(\mu, b) + \varepsilon \cdot (-x)] + \omega^c[(1 - \varepsilon)(\Pi_\mu(\mu, b) + k) + \varepsilon \cdot 0 - k] + \omega^a[(1 - \varepsilon)(\Pi_\mu(\mu, b) + k) + \varepsilon \cdot (-x) - k] = (1 - \varepsilon)\Pi_\mu(\mu, b) - \varepsilon(1 - \omega^c)x - \varepsilon(\omega^c + \omega^a)k$ . Substituting  $\Pi_\mu(\mu, b) = 1 - \omega^a(1 + x)$  into the last expression, we derive  $\Pi_\mu(\tilde{\mu}, \tilde{b}) = 1 - \omega^a(1 + x) - \varepsilon[1 - \omega^a(1 + x) + (1 - \omega^c)x + (\omega^c + \omega^a)k]$ .

Since  $(\mu, b)$  is an NSC, it must be proof against such a mutation, that is,  $\exists \bar{\varepsilon} > 0$  s.t.  $\forall \varepsilon < \bar{\varepsilon} \quad \Pi_{\theta^*}(\tilde{\mu}, \tilde{b}) \leq \Pi_\mu(\tilde{\mu}, \tilde{b})$ , that is,  $(1 - \varepsilon)(1 - \omega^c)y \leq 1 - \omega^a(1 + x) - \varepsilon[1 - \omega^a(1 + x) + (1 - \omega^c)x + (\omega^c + \omega^a)k]$ , or equivalently,  $(1 - \omega^c)y - 1 + \omega^a(1 + x) + \varepsilon[1 - \omega^a(1 + x) + (1 - \omega^c)(x - y) + (\omega^c + \omega^a)k] \leq 0$ . For this inequality to hold for every  $\varepsilon < \bar{\varepsilon}$ , it must be that  $(1 - \omega^c)y - 1 + \omega^a(1 + x) < 0$ , or equivalently,  $\omega^c > \frac{1+x}{y} \cdot \omega^a + \frac{y-1}{y}$ . Thus, condition (ii) must hold.

Finally, cognitive heterogeneity of  $(\mu, b)$  implies  $\tau > 0$ , and hence  $1 > \omega^c + \omega^a > \frac{1+x+y}{y} \cdot \omega^a + \frac{y-1}{y}$ , which is equivalent to  $\omega^a < \frac{1}{1+x+y}$ . Thus, condition (i) must also hold.  $\square$



# Chapter 2

## Idiosyncratic Preferences in Games on Networks

**Abstract:** This paper considers a fixed network of players endowed with idiosyncratic preferences over actions and involved in interactions of various types. The aim is to investigate the interplay between idiosyncratic preferences and interactional incentives on a network. The earlier literature demonstrated the conflict between players' intrinsic preferences and coordination incentives. This paper shows that such a conflict is also present in contexts in which players do not necessarily aim at coordination with their peers. The introduction of action preferences changes equilibrium outcomes in a non-trivial fashion: some equilibria disappear, while other, qualitatively new ones, appear. We characterize equilibria for a large class of games, including games of strategic complements and strategic substitutes, and outline a subclass in which following idiosyncratic action preferences is a unique equilibrium. This equilibrium is Pareto optimal and for many games is also a unique efficient profile.

**JEL codes:** C62, C72, D85.

**Keywords:** *network games; network effects; idiosyncratic preferences; preference heterogeneity; efficiency.*

### 2.1 Introduction

It is quite common that decisions of an individual are influenced not only by her intrinsic, idiosyncratic preferences over alternatives but also by analogous decisions of her business or personal contacts. While the first decision factor is classical in economic theory, the importance of the second one was recognized more recently and modeled within one of the branches of network economics – games on networks – that extends game-theoretic reasoning to a network setting.

In the last decade the networks literature has undertaken an in-depth study of how social networks influence individuals' decisions.<sup>1</sup> In most of the literature, the only dimension of heterogeneity between players is their structural position in the network, since such an approach allows to isolate network effects.<sup>2</sup> This paper considers an additional source of heterogeneity – players' idiosyncratic preferences over available actions, which implies that their utilities have two distinct components. The *idiosyncratic utility component* derives from concordance between a player's idiosyncratic preference and her action choice, while the *interactional utility component* originates from the network position of a player and depends on the nature of a particular underlying game. The aim of this paper is to study the specific interplay between idiosyncratic preferences and various types of interactions on a network.

We introduce idiosyncratic action preferences into a large class of semi-anonymous graphical games on a fixed network in the complete information setting.<sup>3</sup> In particular, we consider binary action games of strategic complements and strategic substitutes.<sup>4</sup> Players, endowed with idiosyncratic preferences over actions, make their action choices simultaneously. Semi-anonymity of the payoff functions, coupled with linearity with respect to the number of neighbors choosing each action, result into threshold best response functions. The thresholds are different for players of different degrees, as in semi-anonymous graphical games, but also for players with different idiosyncratic action preferences.<sup>5</sup> We aim to characterize corresponding equilibrium sets and determine how equilibrium existence and uniqueness depend on various parameters of the model. We compare equilibrium sets in different games with respective efficient (welfare maximizing) action profiles for some standard network structures (stars and complete networks) and then derive several interesting implications for arbitrary networks. We also investigate a special class of *fully satisfying* equilibria – those that maximize overall idiosyncratic utility of players and that might be of interest, for instance, when

---

<sup>1</sup>The latest survey is Bramoullé and Kranton (2016). See also a seminal work of Galeotti et al. (2010) and an extensive survey by Jackson and Zenou (2014).

<sup>2</sup>In the theoretical literature exceptions usually concern models of network formation, which introduce heterogeneity in the cost of interaction (Golub and Livne (2011)) or in benefits from socialization (Currarini et al. (2009), Cabrales et al. (2011)). In empirical applications there are more models with heterogeneity in players' characteristics: see e.g. Calvó-Armengol et al. (2009) for education, or Patacchini and Zenou (2012) for crime.

<sup>3</sup>In *graphical games*, first introduced in Kearns et al. (2001), each player's payoff is affected by just a subset of players – her neighbors in the network. In *semi-anonymous* graphical games all neighbors influence a player's payoff in a symmetric fashion, that is, a player cares only about the total number of her neighbors playing each of the actions and not about who plays what. For formal definitions see Jackson (2008).

<sup>4</sup>In a game of *strategic complements*, a player is more inclined to choose a particular action as more of her neighbors choose it. For *strategic substitutes* the opposite holds: the less chosen by the neighbors the more attractive an action is for a player.

<sup>5</sup>For more on threshold best response functions, see Galeotti et al. (2010) or Jackson (2008).

different utility components are enjoyed by distinct agents (see the school choice example below).

To illustrate our setting, let us consider two choice situations, in which network interactions exhibit, respectively, strategic complements or strategic substitutes.

First, consider a group of schoolchildren in their final year, connected in a friendship network. Each of them has to decide whether to continue education (e.g. enter a university) or to go to the labor market. Two factors influence each one's decision: their idiosyncratic preference over the two options and their interactional preference – to match their choice with as many friends as possible (strategic complements). Now, consider a group of students, again connected in a network, who have to choose between two overlapping informational events (e.g. workshops, study courses, parallel sessions at a conference etc). Each of them has an idiosyncratic preference over the alternatives and at the same time wants more of her peers to choose a different event in order to receive more information about the missed one (strategic substitutes). Note that in both examples the network is formed prior to the action choice, decisions have to be made simultaneously by all players, and complete information – common knowledge of the network and players' preferences – seems to be a relevant assumption for relatively small numbers of connected players. The key feature of both situations, however, is the generic tension between idiosyncratic and interactional (here strategic complements/substitutes) incentives.

As mentioned above, in some contexts these two incentives might even correspond to distinct agents. To illustrate possible disentanglement of these two utility components, let us slightly modify the first example. Consider a class of final year primary school children, whose parents have to make choices about whether their respective kids will go to school A or school B (e.g. public or private secondary school, a gymnasium or a "general" school etc). The kids' friendship network is already formed and known to the parents. Kids have coordination preferences (to go to the same school with their friends), while parents have idiosyncratic preferences over the schooling alternatives. Every parent seeks to maximize her own and her child's utility, while each then enjoys their own part. The fully satisfying action profile in this setting corresponds to parents' maximum welfare, while kids' welfare is maximized if all kids go to the same school, no matter which. What is interesting is how these profiles relate to those maximizing overall social welfare. Our results suggest that if parents' preferences over schools are very strong, then the fully satisfying action profile – when all kids go to the schools preferred by their respective parents – maximizes not only parents' but also overall welfare. On the other hand, if kids' coordination preferences are very strong, then overall social welfare is maximized in one of homogeneous action profiles – namely, when all kids go to the school preferred by the majority of parents. This latter case is quite interesting, as parents' preferences serve here as a kind of selection mechanism for

welfare maximizing profiles.

In general, two utility sources – idiosyncratic and interactional – can interact in different ways, making respective utility components independent or interdependent. To better illustrate these two possibilities, consider the following classical example of a network game with strategic complements. Let a fairly small network of firms or individuals, where connections represent established partnerships, face a binary technology choice. Suppose the two technologies are not perfectly compatible (e.g. Mac and Windows), so that the interactional incentive is to match own technology choice with as many partners as possible.<sup>6</sup> The agents are heterogeneous in the sense that each of them idiosyncratically prefers one or another technology – this might be a hedonic preference, a monetary-driven incentive, a status quo anchor etc, but for the purpose of our illustration let us assume that these preferences originate from idiosyncratic costs of technology adoption. If the technology cost is a lump sum, the idiosyncratic utility bonus for choosing the preferred alternative is *independent* of the network and, in particular, of a player’s degree of connectivity. However, if the cost depends on the intensity of usage (is paid per connection), then the two utility components are *interdependent* (a more connected player gets a higher utility bonus in the case of satisfying her idiosyncratic preference than a less connected player). In this paper we consider both possibilities and outline the main differences in the alternative models’ predictions.

It should be justly mentioned that our paper is not the first attempt in the literature on games on networks to account for players’ idiosyncratic preferences. Hernández et al. (2013) introduce preferences over actions in the binary action setting for two specific games, where utility arises either from coordination or from anti-coordination of own action with the neighbors and the size of idiosyncratic utility bonus depends on the network (interdependent relationship between utility components). In the follow-up paper, Hernández et al. (2017), the authors partially characterize and classify equilibria for the case of coordination.<sup>7</sup>

Building on the model proposed in Hernández et al. (2013), we analyze a large class of games, which includes their games as special cases. We consider two different model specifications – with independent and with interdependent relationship between utility components. For both specifications, all games can be grouped into several qualitatively different subclasses, that are characterized by different best response strategies and thus different equilibrium outcomes. Hernández et al. (2017) demonstrated existence of a parameter region with a unique, fully satisfying equilibrium. We show that there exists a whole subclass of

---

<sup>6</sup>Similar interactional incentives arise if the choice has to be made between two mobile network operators (under duopoly) who charge different prices for calls inside and outside the network, or two banks who set lower charges for money transfers between own customers.

<sup>7</sup>Two experimental papers, Ellwardt et al. (2016) and Goyal et al. (2021), seek to test the theoretical predictions of Hernández et al. (2017), adding to the game a network formation stage. However, since our paper models interactions on a fixed network, relevant comparisons to these papers cannot be made.

games, both of strategic complements and of strategic substitutes, where such an equilibrium exists and is unique. Moreover, it is Pareto efficient and, for some games, is even a unique efficient profile.

We also provide necessary and sufficient conditions for existence of a fully satisfying equilibrium in all other games: what is decisive is sufficient segregation (for strategic complements) or interconnection (for strategic substitutes) of two preference groups of players. For our school choice example it would mean that if kids' coordination preferences are strong, then all parents send their kids to their preferred schools if and only if all kids have sufficiently many friends whose parents prefer the same school.

For existence of other (not necessarily fully satisfying) equilibria with heterogeneity in action, necessary and sufficient conditions are similar: existence of sufficiently segregated (or interconnected) subsets of players. However, it is no longer required that these subsets of players correspond to different preference groups. In our school choice example, if kids' coordination preferences are strong and their friendship network consists of two sufficiently segregated groups, each including kids whose parents prefer different schools, then it is possible that all kids end up studying together with most of their friends, while only some parents satisfy their school preferences.

The rest of the paper is organized as follows. Section 2 presents the model. Sections 3 and 4 deal with interdependent utility specification and present, respectively, players' best response functions and equilibrium analysis for different classes of games. The last subsection of the equilibrium analysis illustrates the results for standard network structures and compares equilibrium and efficient action profiles. Section 5 discusses the impact of idiosyncratic preferences on equilibrium outcomes and outlines the main differences between independent and interdependent utility specifications. Section 6 briefly concludes and appendix contains proofs of the results.

## 2.2 The model

### 2.2.1 Games and idiosyncratic action preferences

Let  $G$  be a network (graph) with the set of nodes  $N = \{1, \dots, n\}$  and links represented by an adjacency matrix. We consider undirected unweighted networks, i.e. the adjacency matrix is symmetric with entries  $G_{ij} \in \{0, 1\}$  for all  $i, j \in N$  (with 1 implying a link between  $i$  and  $j$ , and 0 implying no link). By convention,  $G_{ii} = 0$  for all  $i \in N$ .<sup>8</sup> For a node  $i$  we denote the set of  $i$ 's neighbors in  $G$  by  $N_i(G) = \{j \in N \mid G_{ij} = 1\}$  and the cardinality of this set, called

---

<sup>8</sup>All our results qualitatively hold under an alternative assumption,  $G_{ii} = 1$  for all  $i \in N$ , which is used in Hernández et al. (2017).

also  $i$ 's *degree*, by  $d_i$ . Given a network  $G$ , we let the set of its nodes be the set of players and  $X = \{0, 1\}$  be the action set, the same for all players.  $x_i \in X$  denotes  $i$ 's action in an action profile  $x = (x_1, \dots, x_n)$ ,  $x_{-i} \in X^{n-1}$  – the vector of actions of all players except  $i$  and  $x_{N_i(G)} \in X^{d_i}$  – the vector of actions of  $i$ 's neighbors in  $G$ .

We assume that each player has a strict idiosyncratic preference over the actions that is exogenous and does not change throughout the game. Obviously, the preference set coincides with the action set:  $\Theta = \{0, 1\}$ . Similar to an action profile, an (idiosyncratic) *preference profile*  $\theta = (\theta_1, \dots, \theta_n)$  is a vector of idiosyncratic preferences of all players in the network. We call a preference profile *homogeneous* if  $\theta_i = \theta_j$  for all  $i, j \in N$ , otherwise we call it *heterogeneous*. We denote by  $N^\pi (\subseteq N)$  the subset of players with preference  $\pi \in \{0, 1\}$ . For a heterogeneous preference profile,  $\{N^0, N^1\}$  partitions the set of players into two *preference groups*, whose respective cardinalities are denoted by  $n^0$  and  $n^1$ . Whenever it does not create confusion with the common terminology, we use the term *network* to refer to a pair  $(G, \theta)$ , combining a network structure and the distribution of idiosyncratic preferences in this network, which is assumed to be common knowledge prior to the game.

The payoff for a player  $i$  with idiosyncratic action preference  $\theta_i$  and the set of neighbors  $N_i(G)$  is defined as follows:

$$u_i(\theta_i, x_i, x_{N_i(G)}) = \sum_{j \in N_i(G)} (\delta \cdot \mathbf{1}_{\{x_i=x_j\}} + (1 - \delta) \cdot \mathbf{1}_{\{x_i \neq x_j\}} + \lambda \cdot \mathbf{1}_{\{x_i=\theta_i\}}) \quad (2.1)$$

where  $\delta \in [0; 1]$  and  $\lambda \in [0; \infty)$ .

The first parameter  $\delta$  reflects relative advantage of matching versus mismatching of own action with the neighbors' actions. If  $\delta > \frac{1}{2}$ , it is a game of strategic complements, if  $\delta < \frac{1}{2}$  – a game of strategic substitutes. The second parameter  $\lambda$  determines a utility bonus that a player gets for each of her connections if she chooses her preferred action. Thus,  $\lambda$  reflects the strength of idiosyncratic preferences: the higher  $\lambda$ , the stronger idiosyncratic preferences and the larger utility loss if a player cannot choose her preferred action. Since we are interested in the impact of introducing exogenous heterogeneity between players, in the rest of the paper we consider  $\lambda > 0$ . The case  $\lambda = 0$  corresponds to the framework without idiosyncratic action preferences and the difference between the two frameworks is discussed in section 2.5.

To better illustrate the nature of the game, let us consider an (isolated) pair of connected players. Tables 2.1 and 2.2 represent their incentives in normal form for the case when players have the same idiosyncratic action preferences and for the case when their preferences differ.

	<b>0</b>	<b>1</b>
<b>0</b>	$\delta + \lambda, \delta + \lambda$	$1 - \delta + \lambda, 1 - \delta$
<b>1</b>	$1 - \delta, 1 - \delta + \lambda$	$\delta, \delta$

Table 2.1: The game between two players with the same idiosyncratic action preference:  $\theta_i = \theta_j = 0$

	<b>0</b>	<b>1</b>
<b>0</b>	$\delta + \lambda, \delta$	$1 - \delta + \lambda, 1 - \delta + \lambda$
<b>1</b>	$1 - \delta, 1 - \delta$	$\delta, \delta + \lambda$

Table 2.2: The game between two players with different idiosyncratic action preferences:  $\theta_i = 0$  for the row player and  $\theta_j = 1$  for the column player

Note that in the utility function (2.1) player  $i$ 's idiosyncratic utility bonus  $\lambda \cdot d_i$  depends on the network. In section 2.5 we discuss an alternative utility function, in which idiosyncratic and interactional components are additively separable:

$$u_i(\theta_i, x_i, x_{N_i(G)}) = \sum_{j \in N_i(G)} (\delta \cdot \mathbb{1}_{\{x_i = x_j\}} + (1 - \delta) \cdot \mathbb{1}_{\{x_i \neq x_j\}}) + \lambda \cdot \mathbb{1}_{\{x_i = \theta_i\}}. \quad (2.2)$$

As a final remark, let us point out that the games we consider in this paper belong to the class of graphical games, and thus, without loss of generality, we can assume that  $G$  is a connected network, that is, every two nodes are connected by some path in  $G$ . If the network is disconnected, each of its components can be analyzed separately and all the results of this paper will hold componentwise.

## 2.2.2 Equilibrium concept

We consider a complete information setting with rational players. All players, given the network  $(G, \theta)$ , simultaneously choose actions that maximize their respective payoffs. For each given network we analyze a class of games  $\Gamma = \{\Gamma_{\delta, \lambda} \mid 0 \leq \delta \leq 1, \lambda > 0\}$ , where every specific game is determined by two parameters  $\delta$  and  $\lambda$  (with a slight abuse of terminology, we will refer to a pair  $(\delta, \lambda)$  as "a game", implying the corresponding  $\Gamma_{\delta, \lambda}$ ). The equilibrium concept is based on the  $n$ -player Nash for a fixed network. We refine it by excluding set-valued best responses by means of a quite natural tie-breaking rule: in the case of payoff indifference a player always chooses her preferred action. This implies, in particular, that all equilibria are pure strategy equilibria.

**Definition 1.** For a game  $\Gamma_{\delta, \lambda}$  on a network  $(G, \theta)$ , an action profile  $x = (x_1, \dots, x_n)$  is a *selfish Nash equilibrium (SNE)* if it is a Nash equilibrium and for all players  $i \in N$  the following holds: if  $\exists x'_i \neq x_i$  s.t.  $u_i(\theta_i, x'_i, x_{N_i(G)}) = u_i(\theta_i, x_i, x_{N_i(G)})$ , then  $x_i = \theta_i$ .

We differentiate between *homogeneous equilibria*, in which all players choose the same action, and *heterogeneous*, in which both actions are chosen. What also matters for comparison of equilibria in our framework is whether players are able to satisfy their idiosyncratic action preferences. Following the terminology in Hernández et al. (2013), we call an action  $x_i$  *satisfying* for player  $i$  if  $x_i = \theta_i$ , otherwise we call it *frustrating*. A player who chooses a satisfying (frustrating) action in  $x$  is called a *satisfied (frustrated)* player.

**Definition 2.** For a game  $\Gamma_{\delta,\lambda}$  on a network  $(G, \theta)$ , an action profile  $x$  is called *fully satisfying* if  $x_i = \theta_i \forall i \in N$ . If a fully satisfying action profile constitutes an equilibrium, it is called a *fully satisfying equilibrium*.

The last two definitions, that will prove useful in our analysis, characterize  $i$ 's neighbors with respect to whether they choose  $i$ 's preferred action. For a player  $i$  with a preferred action  $\theta_i$ , her neighbor  $j$  is called  $i$ 's *companion* in  $x$  if  $x_j = \theta_i$ , otherwise  $j$  is called  $i$ 's *opponent*.

## 2.3 Best response functions

Since a player's payoff depends on her neighbors' actions in an anonymous way, what matters for her decision is just the total number of her neighbors choosing each of the actions. Without loss of generality, let  $\tau_i$  be the number of  $i$ 's neighbors who choose action 1. Due to linearity of payoff functions with respect to  $\tau_i$ , players' best responses take the form of threshold functions. For a given game, best response functions for all players of the same degree and with the same action preference coincide.

It appears that the whole parameter space  $(\delta, \lambda)$ , which represents the range of games  $\Gamma$ , splits into three regions that correspond to qualitatively different behaviors, and eventually to different equilibria. These are coordination games region  $R_C = \{(\delta, \lambda) \mid \frac{1}{2} < \delta \leq 1, 0 < \lambda < 2\delta - 1\}$ , anti-coordination games region  $R_A = \{(\delta, \lambda) \mid 0 \leq \delta < \frac{1}{2}, 0 < \lambda < 1 - 2\delta\}$  and the in-between region corresponding to dominant action games  $R_D = \{(\delta, \lambda) \mid 0 \leq \delta \leq 1, \lambda \geq |2\delta - 1|\}$ . We analyze these three classes of games in turn.

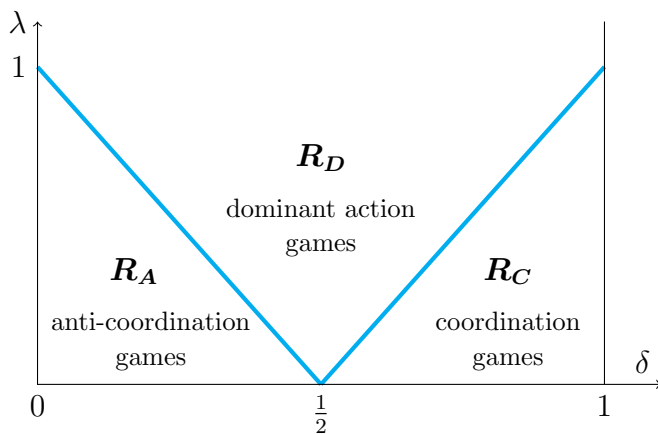


Figure 2.1: Three parameter regions, representing three classes of games.



### 2.3.1 Coordination games

Before we characterize players' best responses, let us define special partitions of  $R_C$  consisting of  $L$  parts, where  $L = \lceil \frac{d_i}{2} \rceil$ .<sup>9,10</sup> As we will see, within each of these parts (subregions) the best response functions of all players of the same degree and with the same action preference are identical.

For a player  $i$  of degree  $d_i$  the partition  $\{R_C^1(d_i), \dots, R_C^L(d_i)\}$  is defined as follows (for a graphical illustration see Figure 2.2):

$$R_C^l(d_i) = \{(\delta, \lambda) \in R_C : \frac{(1 + \lambda)d_i - 2(l - 1)}{2(d_i - 2(l - 1))} < \delta \leq \frac{(1 + \lambda)d_i - 2l}{2(d_i - 2l)}\}$$

for  $l = 1, \dots, L - 1$ , and

$$R_C^L(d_i) = \{(\delta, \lambda) \in R_C : \frac{(1 + \lambda)d_i - 2(L - 1)}{2(d_i - 2(L - 1))} < \delta \leq 1\}.$$

#### **Proposition 1** [BEST RESPONSES. COORDINATION GAMES]

*In a game  $(\delta, \lambda) \in R_C$ , the best response function of a player  $i$  with preference  $\theta_i$  and  $d_i$  neighbors,  $\tau_i$  of whom play 1, is*

$$BR_i(\theta_i, d_i, \tau_i) = \begin{cases} 1, & \text{if } \tau_i > \tau_{\delta, \lambda}^{\theta_i}(d_i) \\ 0, & \text{if } \tau_i < \tau_{\delta, \lambda}^{\theta_i}(d_i) \\ \theta_i, & \text{if } \tau_i = \tau_{\delta, \lambda}^{\theta_i}(d_i) \end{cases}$$

where  $\tau_{\delta, \lambda}^{\theta_i}(d_i) = \theta_i l + (1 - \theta_i)(d_i - l)$  for  $(\delta, \lambda) \in R_C^l(d_i)$ ,  $l = 1, \dots, L$ .

To visualize this result, consider the following figure depicting subregions of parameter values corresponding to different thresholds. The upper subregion  $R_C^1(d_i)$  covers the cases with high strength of idiosyncratic preferences relative to coordination incentives: here having one companion already suffices for a player to choose her preferred action. The weaker idiosyncratic preferences or stronger coordination incentives, the more companions a player needs in order to follow her action preference. Thus, for a given degree  $d_i$ , the need for companions monotonically increases with  $\delta$  and decreases with  $\lambda$ , and the maximum possible companion requirement is  $\lceil \frac{d_i}{2} \rceil$  – the majority of neighbors.<sup>11</sup>

<sup>9</sup>Here  $\lceil x \rceil$  denotes the ceiling of  $x$ .

<sup>10</sup>Although  $L$  is a function of degree  $d_i$ , we omit the argument whenever it does not create confusion, in order to avoid cumbersome notation.

<sup>11</sup>If the number of companions exceeds  $\lceil \frac{d_i}{2} \rceil$ , there is no longer conflict between idiosyncratic preferences and interactional incentives.

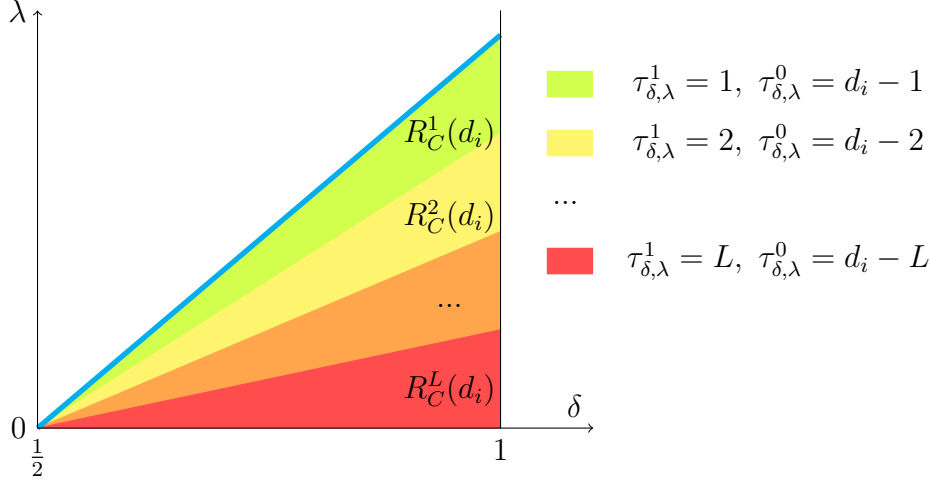


Figure 2.2: Decision thresholds for a player of degree  $d_i$  for different coordination games.

Note that if none of  $i$ 's neighbors chooses her preferred action,  $i$  will not choose it either. In other words, in coordination games every player needs at least one companion in order to follow her idiosyncratic action preference.

**Corollary 1** (MINIMUM COMPANION REQUIREMENT). *In a game  $(\delta, \lambda) \in R_C$ ,*

$$\forall i \in N : BR_i(\theta_i, d_i, \tau_i) = \theta_i \Rightarrow \exists j \in N_i(G) \text{ s.t. } x_j = \theta_i.$$

It is also quite straightforward to prove that the companion requirement is weakly increasing in degree.

**Corollary 2** (ADDING/DELETING A LINK). *For all degrees  $d_i \geq 2$  and  $l = 1, \dots, L$  the following holds:*

$$R_C^l(d_i) \subseteq R_C^l(d_i + 1) \cup R_C^{l+1}(d_i + 1) \quad \text{and} \quad R_C^l(d_i) \subseteq R_C^l(d_i - 1) \cup R_C^{l-1}(d_i - 1),$$

where  $R_C^l(d_i) = \emptyset$  whenever  $l \notin \{1, \dots, L\}$ .

That is, a player of degree  $d_i$  who needs  $l$  companions in order to play her preferred action, would increase this companion requirement by at most one if she gets an additional link, and would decrease this requirement by at most one if she loses one existing link.

### 2.3.2 Anti-coordination games

Similarly to the previous case, let us define the following partition of  $R_A$  for a given degree  $d_i$  (see Figure 2.3):  $\{R_A^1(d_i), \dots, R_A^L(d_i)\}$ , where

$$R_A^l(d_i) = \{(\delta, \lambda) \in R_A : \frac{(1-\lambda)d_i - 2l}{2(d_i - 2l)} \leq \delta < \frac{(1-\lambda)d_i - 2(l-1)}{2(d_i - 2(l-1))}\}$$

for  $l = 1, \dots, L - 1$ , and

$$R_A^L(d_i) = \{(\delta, \lambda) \in R_A : 0 \leq \delta < \frac{(1 - \lambda)d_i - 2(L - 1)}{2(d_i - 2(L - 1))}\}.$$

**Proposition 2** [BEST RESPONSES. ANTI-COORDINATION GAMES]

In a game  $(\delta, \lambda) \in R_A$ , the best response function of a player  $i$  with preference  $\theta_i$  and  $d_i$  neighbors,  $\tau_i$  of whom play 1, is

$$BR_i(\theta_i, d_i, \tau_i) = \begin{cases} 1, & \text{if } \tau_i < \tau_{\delta, \lambda}^{\theta_i}(d_i) \\ 0, & \text{if } \tau_i > \tau_{\delta, \lambda}^{\theta_i}(d_i) \\ \theta_i, & \text{if } \tau_i = \tau_{\delta, \lambda}^{\theta_i}(d_i) \end{cases}$$

where  $\tau_{\delta, \lambda}^{\theta_i}(d_i) = \theta_i(d_i - l) + (1 - \theta_i)l$  for  $(\delta, \lambda) \in R_A^l(d_i)$ ,  $l = 1, \dots, L$ .

Note that subregions corresponding to different thresholds in the case of anti-coordination (Figure 2.3) are symmetric to the corresponding subregions for coordination games (Figure 2.2). If idiosyncratic preferences are strong compared to anti-coordination incentives (subregion  $R_A^1(d_i)$ ), having one opponent is sufficient to play the preferred action. The minimal number of such opponents increases with anti-coordination incentives and with weaker idiosyncratic preferences until it reaches its maximum possible value  $\lceil \frac{d_i}{2} \rceil$ .

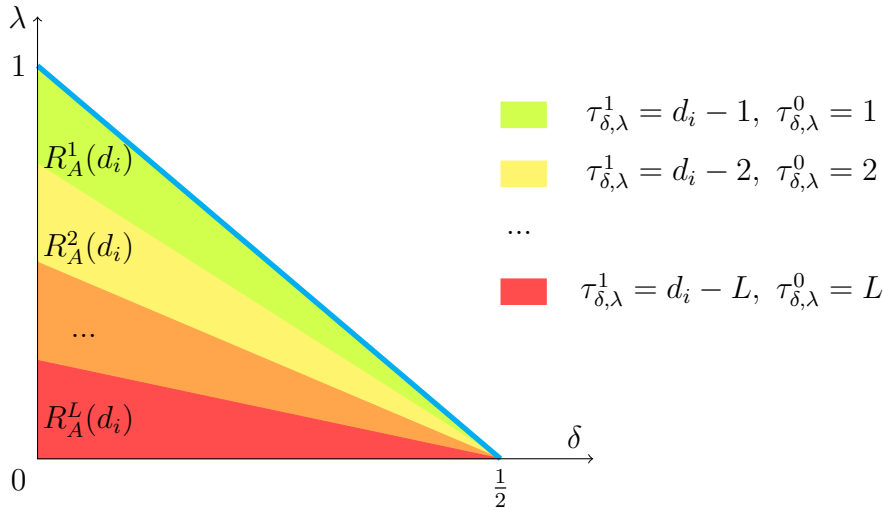


Figure 2.3: Decision thresholds for a player of degree  $d_i$  for different anti-coordination games.

Similar to the coordination case, the following corollary provides a necessary condition for choosing the preferred action in an anti-coordination game: a player needs at least one opponent.

**Corollary 3** (MINIMUM OPPONENT REQUIREMENT). *In a game  $(\delta, \lambda) \in R_A$ ,*

$$\forall i \in N : BR_i(\theta_i, d_i, \tau_i) = \theta_i \Rightarrow \exists j \in N_i(G) \text{ s.t. } x_j = 1 - \theta_i.$$

The opponent requirement is weakly increasing in degree. Similar to the coordination case, a player with  $d_i$  links who needs  $l$  opponents would increase her opponents requirement by at most one if she gets an additional link, and would decrease this requirement by at most one if she loses one existing link.

### 2.3.3 Dominant action games

The in-between region  $R_D$  contains both games of strategic complements and those of strategic substitutes. However, these are the games in which neither coordination nor anti-coordination incentives are well-pronounced. It appears that for this class of games the unique individually rational strategy is to follow idiosyncratic action preferences.

The intuition behind this result is simple: when idiosyncratic preferences become relatively more important than interactional incentives (the idiosyncratic utility bonus outweighs the utility difference between matching and mismatching a neighbor's action), players no longer take interactional incentives into account and choose exclusively according to their idiosyncratic preferences. Acting in such a way becomes a strictly dominant strategy for every player.

**Proposition 3** [BEST RESPONSES. DOMINANT ACTION GAMES]

*In a game  $(\delta, \lambda) \in R_D$ , the best response for every player is her idiosyncratically preferred action:*

$$BR_i(\theta_i, d_i, \tau_i) = \theta_i \quad \forall i \in N.$$

*Proof.* A player  $i$  with preference  $\theta_i \in \{0, 1\}$  has two possible actions:  $\theta_i$  or  $1 - \theta_i$ . The utility gain she gets from each her connection depends on her own action (rows) and that of a corresponding neighbor (columns):

	$\theta_i$	$1 - \theta_i$
$\theta_i$	$\delta + \lambda$	$1 - \delta + \lambda$
$1 - \theta_i$	$1 - \delta$	$\delta$

If  $\delta \in (\frac{1}{2}; \frac{1+\lambda}{2}]$  then  $1 - \delta < \delta \leq 1 - \delta + \lambda < \delta + \lambda$ . This means that strategy  $\theta_i$  is strictly dominant for player  $i$  (in the case of indifference, the tie-breaking rule applies). If  $\delta \in [\frac{1-\lambda}{2}; \frac{1}{2})$  then  $\delta < 1 - \delta \leq \delta + \lambda < 1 - \delta + \lambda$ . Hence, strategy  $\theta_i$  is strictly dominant for  $i$  also in this case. Finally, if  $\delta = \frac{1}{2}$  then strict dominance of  $\theta_i$  follows from the fact that  $\lambda > 0$ . As the above is true for every  $i$ 's neighbor in  $G$ , we can conclude that  $\theta_i$  is player  $i$ 's unique best response.  $\square$

### 2.3.4 Companion/opponent requirement

Let us summarize here players' best response behavior. For this purpose, let us define a function  $l : [0, 1] \times (0; +\infty) \times \mathbb{N} \rightarrow \mathbb{N} \cup \{0\}$  that maps every game  $(\delta, \lambda)$  and every possible degree  $d_i \in \mathbb{N}$  of a player to the minimum number of companions (for  $\delta \geq \frac{1}{2}$ ) or opponents (for  $\delta \leq \frac{1}{2}$ ) a player of this degree needs in order to play her preferred action. That is, the value of the function  $l$  corresponds to a natural number labelling the corresponding subregion  $R_C^l(d_i)$  or  $R_A^l(d_i)$  (or to zero for  $R_D$ ).

For analysis of equilibria in different games it is convenient to consider  $\delta$  and  $\lambda$  as parameters and use the function  $l$  as a function of a single argument – a player's degree:  $l(\delta, \lambda, d_i) = l_{\delta, \lambda}(d_i)$ . The following lemma allows to derive  $l$  for a given game  $(\delta, \lambda)$  and a player's degree  $d_i$ .

**Lemma 1** (COMPANION/OPPONENT REQUIREMENT).

(i) In a game  $(\delta, \lambda) \in R_C$  ( $R_A$ ), the minimum number of companions (opponents) that a player  $i$  of degree  $d_i$  needs in order to play her preferred action  $\theta_i$  equals

$$l_{\delta, \lambda}(d_i) = l^* + \mathbf{1}_{\lambda < \tilde{\lambda}(l^*)},$$

where  $l^* = \operatorname{argmin}_{m=1 \dots L} |\lambda - \tilde{\lambda}(m)|$  and  $\tilde{\lambda}(m) = \frac{|2\delta - 1| \cdot (d_i - 2m)}{d_i}$ .<sup>12,13</sup>

(ii) In a game  $(\delta, \lambda) \in R_D$ , a player  $i$  does not need any companions or opponents in order to play her preferred action:  $l_{\delta, \lambda}(d_i) = 0$ .

## 2.4 Equilibrium analysis

Using the best response functions derived in the previous section, we seek to characterize the set of selfish Nash equilibria on a given network  $(G, \theta)$  for different classes of games. The first three subsections deal with existence and uniqueness of different types of equilibria for an arbitrary network for coordination, anti-coordination and dominant action games respectively. The last subsection illustrates the results for several standard network structures and discusses the relationship between efficient (welfare maximizing) action profiles and selfish Nash equilibria.

<sup>12</sup>Here  $|x|$  denotes the absolute value of  $x$ .

<sup>13</sup>Note that, for a given degree  $d_i$ , the curve  $\tilde{\lambda}(m)$  separates subregion  $R_C^m$  from  $R_C^{m+1}$  and subregion  $R_A^m$  from  $R_A^{m+1}$ .

## 2.4.1 Coordination games

We can now characterize the set of selfish Nash equilibria in coordination games, given an arbitrary network  $(G, \theta)$ . The following theorem provides existence conditions separately for homogeneous and heterogeneous equilibria. As we will see, there always exist at least two (homogeneous) equilibria for this class of games, thus equilibrium multiplicity is unavoidable. At the same time, existence of a heterogeneous equilibrium is not always guaranteed (counterexample – a star network).

**Theorem 1** [EQUILIBRIA. COORDINATION GAMES]

For a network  $(G, \theta)$  and game  $(\delta, \lambda) \in R_C$ :

(i) two homogeneous equilibria exist,

(ii) a heterogeneous equilibrium exists iff there exists such a partition  $\{S^0, S^1\}$  of  $N$  that the following conditions are satisfied for  $\pi = 0, 1$ :

- $\forall i \in S^\pi \cap N^\pi : |N_i(G) \cap S^\pi| \geq l_{\delta, \lambda}(d_i)$  and
- $\forall i \in S^\pi \cap N^{1-\pi} : |N_i(G) \cap S^\pi| \geq d_i - l_{\delta, \lambda}(d_i) + 1$ .<sup>14</sup>

The first conclusion of the theorem, namely that both homogeneous action profiles are equilibrium profiles, is not very surprising: the same result holds for games of strategic complements without idiosyncratic action preferences.<sup>15</sup> Theorem 1 confirms this result in an extended setting. The second part of the theorem states that existence of a heterogeneous equilibrium is equivalent to existence of a partition  $\{S^0, S^1\}$  of players satisfying several interconnectivity conditions. If we interpret this partition as the partition of players by chosen action and consider its refinement by the partition of  $N$  into preference groups –  $\{S^0 \cap N^0, S^0 \cap N^1, S^1 \cap N^0, S^1 \cap N^1\}$  (some of subsets can be empty) – then the conditions of the theorem guarantee that chosen actions are best responses for all satisfied ( $i \in S^\pi \cap N^\pi$ ) as well as for all frustrated ( $i \in S^\pi \cap N^{1-\pi}$ ) players.

Hence, Theorem 1 provides an algorithm for practical derivation of all heterogeneous equilibria for any given network by enumerating all possible partitions of the set of players into two subsets and checking whether they satisfy well-defined connectivity conditions. Since the conditions are necessary and sufficient, the number of heterogeneous equilibria is given by the number of such partitions.

Necessary and sufficient conditions for existence of heterogeneous equilibria can be reformulated in cohesion terminology.<sup>16</sup> For this purpose, we define the degree partition of a network and a  $(r_1, \dots, r_K)$ -cohesive partition of a subset of nodes.

<sup>14</sup>Here  $|S|$  denotes the cardinality of a set  $S$ .

<sup>15</sup>See, for instance, Galeotti et al. (2010).

<sup>16</sup>See Morris (2000), as well as chapter 9.6 in Jackson (2008).

**Definition 3** (Mahadev and Peled, 1995). Let  $G$  be a network with distinct positive degrees  $d_{(1)} < \dots < d_{(M)}$ . Define  $D_m = \{i \in N \mid d_i = d_{(m)}\}$  for  $m = 1, \dots, M$ . Then the set-valued vector  $D(G) = (D_1, \dots, D_M)$  is called the *degree partition* of  $G$ .

**Definition 4.** A partition  $\{S_1, \dots, S_K\}$  of a subset of nodes  $S \subset N$  in a network  $G$  is  $(r_1, \dots, r_K)$ -*cohesive* if for  $k = 1, \dots, K$ :

$$\min_{i \in S_k} \frac{|N_i(G) \cap S|}{|N_i(G)|} \geq r_k.$$

That is, a partition of a subset is  $(r_1, \dots, r_K)$ -cohesive if the share of inward-looking links (with nodes from the same subset) for every node from respective  $S_k$  is at least  $r_k$ .

Given  $(G, \theta)$ , we now refine an arbitrary partition  $\{S^0, S^1\}$  of  $N$  using the degree partition of  $G$  and the preference partition  $\{N^0, N^1\}$ .

**Corollary 4** (HETEROGENEOUS EQUILIBRIA. COORDINATION GAMES). *For a network  $(G, \theta)$  with degree partition  $(D_1, \dots, D_M)$  and a game  $(\delta, \lambda) \in R_C$ , a heterogeneous equilibrium exists iff there exists such a partition  $\{S^0, S^1\}$  of  $N$  that for every  $\pi \in \{0, 1\}$  the (possibly trivial) partition  $\{S^\pi \cap N^\pi \cap D_1, \dots, S^\pi \cap N^\pi \cap D_M, S^\pi \cap N^{1-\pi} \cap D_1, \dots, S^\pi \cap N^{1-\pi} \cap D_M\}$  of  $S^\pi$  is  $\left(\frac{l_{\delta, \lambda}(d_{(1)})}{d_{(1)}}, \dots, \frac{l_{\delta, \lambda}(d_{(M)})}{d_{(M)}}, 1 - \frac{l_{\delta, \lambda}(d_{(1)})-1}{d_{(1)}}, \dots, 1 - \frac{l_{\delta, \lambda}(d_{(M)})-1}{d_{(M)}}\right)$ -cohesive.<sup>17</sup>*

This formulation provides additional intuition for Theorem 1: *for maintaining variation in behavior in a network where players have coordination incentives, it is important that there exist two groups of players with sufficient interconnection within each group.* Importantly, it does not matter for existence of a heterogeneous equilibrium *whether these groups coincide with preference groups or not*: sufficient connectivity within groups guarantees that respective action choices are best responses for both satisfied and frustrated players. Figure 2.8(b) in section 2.5 provides an example of a heterogeneous equilibrium in a coordination game that does not coincide with the preference profile.

Next, we provide necessary and sufficient conditions for existence of a fully satisfying equilibrium as such that guarantees the highest degree of satisfaction of idiosyncratic preferences in a network.

**Theorem 2** [EXISTENCE OF A FULLY SATISFYING EQUILIBRIUM. COORDINATION GAMES] *For a network  $(G, \theta)$  and game  $(\delta, \lambda) \in R_C$ , a fully satisfying equilibrium exists iff the following holds:*

$$\forall i \in N : |N_i(G) \cap N^{\theta_i}| \geq l_{\delta, \lambda}(d_i).$$

---

<sup>17</sup>Here a *trivial* partition is such that contains empty subsets.

That is, for coordination games such an equilibrium exists if and only if every player has at least  $l_{\delta,\lambda}(d_i)$  distinct neighbors whose action preferences coincide with her own. Given a network and a preference profile, this condition is very easy to check.

Let us consider an example. Figures 2.4(a) and 2.4(b) depict the same network structure with two different preference profiles. Players' action preferences are denoted by coloured numbers outside circles, while numbers inside circles identify players. In network (a) the fully satisfying action profile constitutes an equilibrium, since every player has sufficiently many neighbors with the same action preference. In network (b) the fully satisfying action profile is not an equilibrium, since conditions of Theorem 2 for player 5 (and if  $l_{\delta,\lambda}(4) = 2$ , also for player 3) are not satisfied.

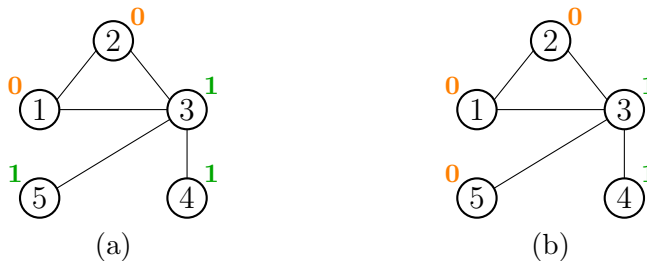


Figure 2.4: Coordination games. A fully satisfying equilibrium exists in (a), but not in (b).

Thus, the distribution of idiosyncratic action preferences on a network is crucial for existence of a fully satisfying equilibrium. But even prior to the preference distribution, some minimal preconditions regarding the size of preference groups must be satisfied: each preference group must include at least two players, otherwise the minimum companion requirement (Corollary 1) will not be satisfied.

On the other hand, a network structure itself might already be decisive. There are such network structures in which fully satisfying equilibria never exist for coordination games, regardless of a preference profile and even of the strength of idiosyncratic preferences. An example of such a network structure is a star network: it never allows for a fully satisfying equilibrium if players have heterogeneous preferences over actions.

## 2.4.2 Anti-coordination games

The following theorem characterizes the set of selfish Nash equilibria for anti-coordination games.

**Theorem 3** [EQUILIBRIA. ANTI-COORDINATION GAMES]

For a network  $(G, \theta)$  and game  $(\delta, \lambda) \in R_A$ :

(i) no homogeneous equilibria exist,



(ii) a heterogeneous equilibrium exists iff there exists such a partition  $\{S^0, S^1\}$  of  $N$  that the following conditions are satisfied for  $\pi = 0, 1$ :

- $\forall i \in S^\pi \cap N^\pi : |N_i(G) \cap S^{1-\pi}| \geq l_{\delta, \lambda}(d_i)$  and
- $\forall i \in S^\pi \cap N^{1-\pi} : |N_i(G) \cap S^{1-\pi}| \geq d_i - l_{\delta, \lambda}(d_i) + 1$ .

Again, the first result, concerning non-existence of homogeneous equilibria, goes along with typical conclusions for anti-coordination games.<sup>18</sup> Consequently, equilibrium existence in general is no longer guaranteed, while in some networks multiplicity of equilibria is still an issue. The second result of the theorem is similar to the corresponding result for coordination games: existence of a heterogeneous equilibrium is equivalent to existence of a partition  $\{S^0, S^1\}$  of network nodes that satisfies specific interconnectivity conditions. Theorem 3 provides an algorithm for practical derivation of all selfish Nash equilibria in a given network for anti-coordination games.

For anti-coordination games the necessary and sufficient conditions for existence of a heterogeneous equilibrium can be reformulated using the notion of outwardness of a partition of a subset, closely related to the notion of cohesion.

**Definition 5.** A partition  $\{S_1, \dots, S_K\}$  of a subset of nodes  $S \subset N$  in a network  $G$  is  $(r_1, \dots, r_K)$ -outward if for  $k = 1, \dots, K$ :

$$\min_{i \in S_k} \frac{|N_i(G) \cap (N \setminus S)|}{|N_i(G)|} \geq r_k.$$

That is, for every node the share of outward-looking links must be at least  $r_k$  (equivalently, the share of inward-looking links must be at most  $1 - r_k$ ) if the node belongs to  $S_k$ .

**Corollary 5** (HETEROGENEOUS EQUILIBRIA. ANTI-COORDINATION GAMES). *For a network  $(G, \theta)$  with degree partition  $(D_1, \dots, D_M)$  and a game  $(\delta, \lambda) \in R_A$ , a heterogeneous equilibrium exists iff there exists such a partition  $\{S^0, S^1\}$  of  $N$  that for every  $\pi \in \{0, 1\}$  the (possibly trivial) partition  $\{S^\pi \cap N^\pi \cap D_1, \dots, S^\pi \cap N^\pi \cap D_M, S^\pi \cap N^{1-\pi} \cap D_1, \dots, S^\pi \cap N^{1-\pi} \cap D_M\}$  of  $S^\pi$  is  $\left(\frac{l_{\delta, \lambda}(d_{(1)})}{d_{(1)}}, \dots, \frac{l_{\delta, \lambda}(d_{(M)})}{d_{(M)}}, 1 - \frac{l_{\delta, \lambda}(d_{(1)})-1}{d_{(1)}}, \dots, 1 - \frac{l_{\delta, \lambda}(d_{(M)})-1}{d_{(M)}}\right)$ -outward.*

Thus, for maintaining variation in behavior in a network where players have anti-coordination incentives, it is important that there exist two groups of players with sufficiently high interconnection between the groups. Here, similar to coordination games, these groups do not have to coincide with preference groups: sufficiently high interconnection between the groups compared to interconnection within each group guarantees existence of a heterogeneous equilibrium.

<sup>18</sup>See, in particular, Bramoullé (2007) and Galeotti et al. (2010).

For existence of a fully satisfying equilibrium in an anti-coordination game it is important that players of the same preference group are not concentrated in the same part of a network. As the following theorem suggests, a fully satisfying equilibrium exists if and only if every player has sufficiently many neighbors with a different action preference. In particular, anti-coordination games can never have fully satisfying equilibria if the preference profile is homogeneous.

**Theorem 4** [EXISTENCE OF A FULLY SATISFYING EQUILIBRIUM. ANTI-COORDINATION GAMES]

For a network  $(G, \theta)$  and game  $(\delta, \lambda) \in R_A$ , a fully satisfying equilibrium exists iff the following holds:

$$\forall i \in N : |N_i(G) \cap N^{1-\theta_i}| \geq l_{\delta, \lambda}(d_i).$$

Figure 2.5 illustrates this result. Here again, numbers inside circles identify players and coloured numbers outside circles correspond to players' action preferences. For a preference profile in (a) the fully satisfying equilibrium exists, since every player has the required number of neighbors with a different preference. In (b) the fully satisfying action profile is not an equilibrium, since for player 4 the requirement is not satisfied.

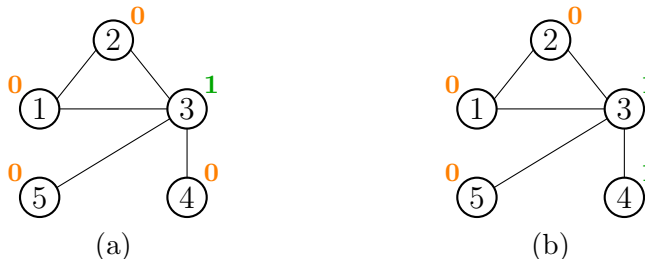


Figure 2.5: Anti-coordination games. A fully satisfying equilibrium exists in (a), but not in (b).

### 2.4.3 Dominant action games

As it follows from Proposition 3, whenever interactional incentives are not well-pronounced, every player chooses her preferred action. The following theorem fully characterizes the set of equilibria for this class of games.

**Theorem 5** [EQUILIBRIA. DOMINANT ACTION GAMES]

For a network  $(G, \theta)$  and game  $(\delta, \lambda) \in R_D$ , there always exists a unique equilibrium – the fully satisfying equilibrium:  $x_i = \theta_i \forall i \in N$ .

The proof is straightforward. Since following own action preference is a unique best response for each player, in equilibrium every player chooses her preferred action. Moreover, since players' preferences over actions are strict, such an equilibrium is unique.

Fully satisfying equilibria maximize idiosyncratic utility of all players. In the case when idiosyncratic and interactional utility components are enjoyed by different groups of agents, fully satisfying action profiles are welfare maximizing for one of the groups. For the other group, welfare maximizing action profiles would be either homogeneous ones (for strategic complements) or those with a maximal number of action mismatches (for strategic substitutes).

As for such action profiles that maximize overall welfare, which we call *efficient* action profiles, they might be neither fully satisfying nor maximizing the overall interactional utility, as we will see, for example, for complete networks (Proposition 7). However, we will also see that there is a non-empty class of games, for which fully satisfying action profiles are not only equilibria but also unique efficient action profiles (Theorem 6).

Note also that in dominant action games every player chooses her preferred action regardless of her neighbors' action choices, thus the equilibrium is a strong Nash, and consequently, Pareto optimal. For these games there exist no Pareto improvements to fully satisfying action profiles.

#### 2.4.4 Efficiency of equilibria

For several standard network structures, namely stars and complete networks, we use the derived results to describe equilibrium sets for the whole range of games and then compare them with respective sets of efficient action profiles. We discuss how efficient action profiles relate to fully satisfying ones and whether and under which conditions they can be achieved as equilibrium outcomes. Insights from the analysis of standard network structures allow to make several conclusions for general network structures.

**Proposition 4** [EQUILIBRIA IN STAR NETWORKS]

*Let  $(G, \theta)$  be a star network with player 1 being the central player.*

- (i) In a game  $(\delta, \lambda) \in R_C$ , an action profile  $x$  is a SNE iff  $x_i = x_j \forall i, j \in N$ .*
- (ii) In a game  $(\delta, \lambda) \in R_A$ , an action profile  $x$  is a SNE iff  $x_i \neq x_1 \forall i \neq 1$ .*
- (iii) In a game  $(\delta, \lambda) \in R_D$ , an action profile  $x$  is a SNE iff  $x_i = \theta_i \forall i \in N$ .*

Any game with strong interactional incentives on a star network has two equilibria: in the case of coordination these are two homogeneous equilibria, in the case of anti-coordination – two heterogeneous, with all peripheral players mismatching the action of the central player. Note that here the equilibrium set is completely independent of a preference profile for all games except for dominant action games.

**Proposition 5** [EFFICIENT ACTION PROFILES IN STAR NETWORKS]

Let  $(G, \theta)$  be a star network with player 1 being the central player.

- (i) For  $\lambda < 2(2\delta - 1)$ , the action profile  $x_i = \theta_1 \forall i \in N$  is efficient. It is a unique efficient action profile unless  $\theta_i = 1 - \theta_1 \forall i \neq 1$ .<sup>19</sup>
- (ii) For  $\lambda < 2(1 - 2\delta)$ , the action profile  $x_i \neq x_1 = \theta_1 \forall i \neq 1$  is efficient. It is a unique efficient action profile unless  $\theta_i = \theta_1 \forall i \in N$ .<sup>20</sup>
- (iii) For  $\lambda > 2 \cdot |2\delta - 1|$ , an action profile  $x$  is efficient iff  $x_i = \theta_i \forall i \in N$ .

Note that in all efficient action profiles the central player satisfies her action preference.<sup>21</sup> This is due to the fact that in a star network the central player has a higher weight in the social welfare function because of her higher connectivity.

Comparison of efficient action profiles with equilibrium action profiles (Figure 2.6) allows us to make several observations.

First, for high values of  $\lambda$  efficient and equilibrium profiles coincide. This is due to the fact that choosing the preferred action gives a very large utility bonus, which outweighs not only the interactional utility difference between two actions for a player (and makes the preferred action dominant) but also the sum of interactional utility differences for a player and all her neighbors (and makes a fully satisfying profile efficient). This result can, in fact, be generalized to an arbitrary network, what we do at the end of this subsection.

Second, for intermediate values of  $\lambda$  efficient action profiles are never attainable in equilibrium. Here idiosyncratic preferences are still strong enough to guide action choices, but overall welfare could have been improved if some players acted against their action preferences (improved through raising interactional utilities of these players' neighbors). Note, however, that the fully satisfying profile is Pareto optimal here (see subsection 2.4.3), thus when overall welfare is improved, some players lose utility.

Third, for low values of  $\lambda$  efficient action profiles are possible in equilibrium but, generically, not guaranteed. For two special cases (see footnotes 19 and 20) the sets of efficient and equilibrium action profiles coincide.

---

<sup>19</sup>If  $\theta_i = 1 - \theta_1 \forall i \neq 1$ , there is one more efficient action profile:  $x_i = 1 - \theta_1 \forall i \in N$ .

<sup>20</sup>If  $\theta_i = \theta_1 \forall i \in N$ , there is one more efficient action profile:  $x_1 \neq x_i = \theta_1 \forall i \neq 1$ .

<sup>21</sup>In two particular cases the central player might be frustrated, but in such cases there always exists another efficient action profile with satisfied central player.

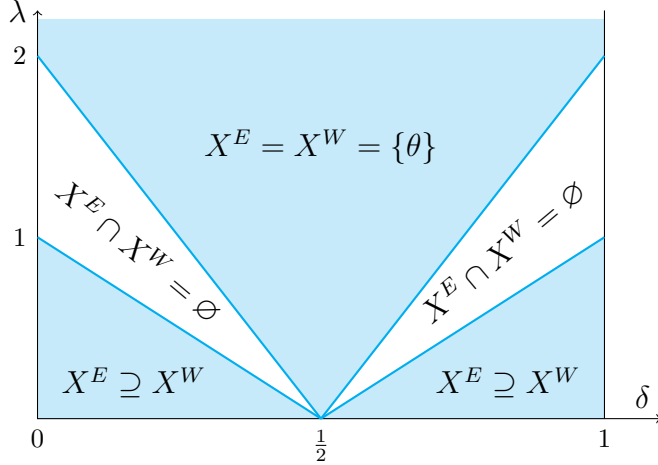


Figure 2.6: Relationship between the equilibrium set  $X^E$  and the set of efficient (welfare maximizing) action profiles  $X^W$  for games on star networks.

Let us turn now to complete networks. Since in a complete network every player has  $n - 1$  neighbors, in the following proposition we use simplified notation  $l := l_{\delta, \lambda}(n - 1)$  for the required number of companions (opponents) for each player in a coordination (anti-coordination) game  $(\delta, \lambda)$ . We additionally denote by  $S^\pi(x) := \{i \in N | x_i = \pi\}$  the subset of players who play action  $\pi \in \{0, 1\}$  in an action profile  $x$ .

**Proposition 6 [EQUILIBRIA IN COMPLETE NETWORKS]**

Let  $(G, \theta)$  be a complete network with  $n$  players.

(i) In a game  $(\delta, \lambda) \in R_C$ , an action profile  $x$  is a SNE iff

- either  $x_i = x_j \forall i, j \in N$ ,
- or  $x_i = \theta_i \forall i \in N$ , if  $n^\pi \geq l + 1 \forall \pi \in \{0, 1\}$ .

(ii) In a game  $(\delta, \lambda) \in R_A$ , an action profile  $x$  is a SNE iff

- either  $x_i = \theta_i \forall i \in N$ , if  $n^\pi \geq l \forall \pi \in \{0, 1\}$ ,
- or  $N^\pi \subset S^\pi(x)$  and  $|S^\pi(x)| = l$ ,
- or  $|S^\pi(x)| = l = \frac{n}{2} \forall \pi \in \{0, 1\}$ .

(iii) In a game  $(\delta, \lambda) \in R_D$ , an action profile  $x$  is a SNE iff  $x_i = \theta_i \forall i \in N$ .

For coordination games the only possible heterogeneous equilibrium is the fully satisfying one. Intuition is quite straightforward: in a complete network there exists no such a partition of players into two groups that players within the same group are more closely connected to each other than to players in the other group, hence departure from action coordination

can only be beneficial if it allows players to satisfy their idiosyncratic preferences. However, such a heterogeneous equilibrium exists only if the preference minority – the smaller of the two preference groups – is sufficiently large and coordination incentives are not too strong ( $l < \lceil \frac{n-1}{2} \rceil$ ). Otherwise only full coordination is possible in equilibrium.

For anti-coordination games heterogeneous equilibria always exist and equilibrium multiplicity is a common issue. It is notable that the preference minority is always satisfied in equilibrium (except for the case of  $l = \frac{n}{2}$ , where any equal split of players between two actions is an equilibrium). If the minority is large, the equilibrium might be unique (and then fully satisfying), while if the minority is smaller, some players from the majority group, in order to make the resulting action profile more balanced, also choose to play the action preferred by the minority. The larger the majority, the more potential players who could take these roles and the more equilibria are possible.

**Proposition 7** [EFFICIENT ACTION PROFILES IN COMPLETE NETWORKS]

Let  $(G, \theta)$  be a complete network with  $n$  players and  $\pi \in \{0, 1\}$  be the action preferred by the minority of players, i.e.  $n^\pi \leq \frac{n}{2}$ .

(i) For  $\lambda < 2(2\delta - 1)\frac{n-n^\pi}{n-1}$ ,  $x$  is efficient iff  $x_i = 1 - \pi \forall i \in N$ .

(ii) For  $\lambda < 2(1 - 2\delta)\frac{n-2n^\pi}{n-1}$ ,  $x$  is efficient iff  $N^\pi \subset S^\pi(x)$  and  $|S^\pi(x)| = \left\lfloor \frac{n}{2} - \frac{\lambda(n-1)}{4(1-2\delta)} \right\rfloor$ .<sup>22</sup>

(iii) For  $\lambda > \max\{2(2\delta - 1)\frac{n-n^\pi}{n-1}, 2(1 - 2\delta)\frac{n-2n^\pi}{n-1}\}$ ,  $x$  is efficient iff  $x_i = \theta_i \forall i \in N$ .

Note that the region where a fully satisfying action profile is efficient is not symmetric about  $\delta = \frac{1}{2}$ . This is due to the fact that welfare benefits of a single player's action switch in a complete network are lower for strategic substitutes than for strategic complements.<sup>23</sup> Thus, in the former case it is more likely that a fully satisfying equilibrium is efficient (see Figure 2.7). Note also that for strategic substitutes multiple efficient profiles are possible for the same network; in all of them, similar to equilibria in anti-coordination games, the preferences minority is satisfied.

Let us now compare the sets of efficient action profiles to the corresponding equilibrium sets for complete networks in more detail.

---

<sup>22</sup>Here  $\lfloor x \rfloor$  denotes the nearest integer to  $x$ .

<sup>23</sup>To see this, consider a preference profile with  $k < \frac{n}{2}$  players with  $\theta_i = \pi$ . In interactional terms, welfare benefits from a single action switch equal  $((n - k) - (k - 1)) \cdot |2\delta - 1|$  for strategic complements (when a minority player switches her action) and  $((n - k - 1) - k) \cdot |2\delta - 1|$  for strategic substitutes (when a majority player switches her action). The idiosyncratic part of welfare benefits is the same in both cases (one player's idiosyncratic utility loss  $\lambda(n - 1)$ ).

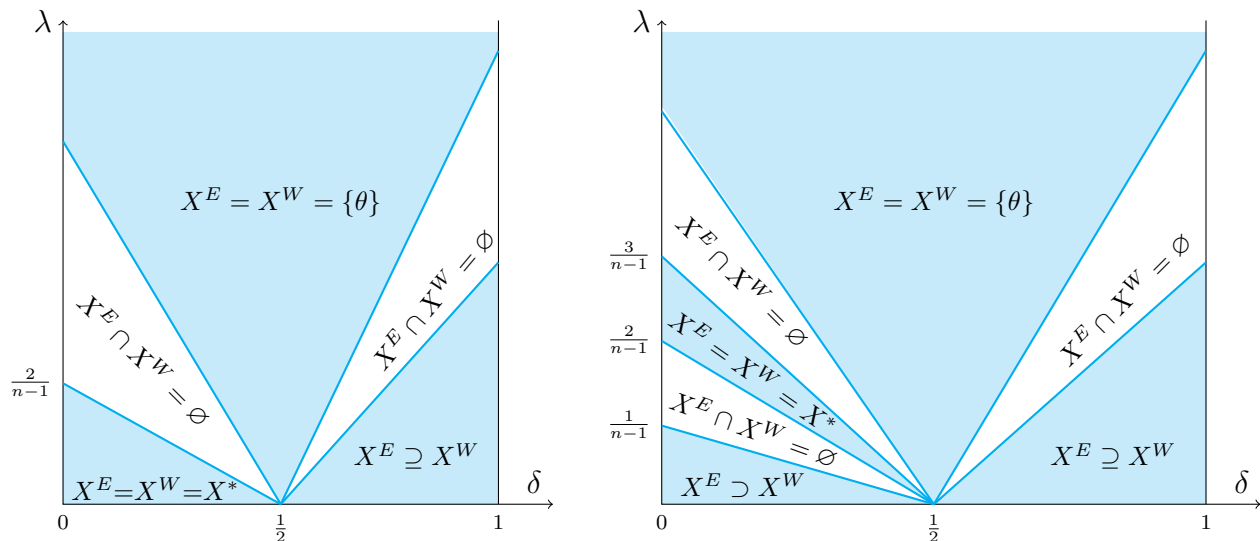


Figure 2.7: Relationship between the equilibrium set  $X^E$  and the set of efficient action profiles  $X^W$  for games on complete networks. Left: for odd  $n$ . Right: for even  $n$ .

For games of strategic complements ( $\delta > \frac{1}{2}$ ) the picture is similar to that for star networks: for high  $\lambda$  the sets coincide and include a single, fully satisfying action profile; for lower  $\lambda$  there is (always non-empty) region of games with no intersection; for low  $\lambda$  a unique (unless  $n^0 = n^1$ ) efficient profile belongs to the equilibrium set.

For games of strategic substitutes ( $\delta < \frac{1}{2}$ ) there are several regions in which the efficient and equilibrium sets coincide at least partially. These are a region with strong idiosyncratic preferences and a region with strong anti-coordination incentives. The latter, where  $X^* = \{x \in X : N^\pi \subset S^\pi(x) \text{ and } |S^\pi(x)| = \lfloor \frac{n-1}{2} \rfloor\}$ , includes a set of profiles with satisfied preferences minority. However, for preference profiles with very large minority ( $n^\pi = \lceil \frac{n-1}{2} \rceil$ ) this region does not exist: not surprisingly, in this case the only efficient profile in any game of strategic substitutes is the fully satisfying one. In a special case – for even number of players and strong anti-coordination incentives ( $l = \frac{n}{2}$ ) – only a subset  $\{x \in X : N^\pi \subset S^\pi(x) \text{ and } |S^0(x)| = |S^1(x)| = \frac{n}{2}\}$  of the equilibrium set  $\{x \in X : |S^0(x)| = |S^1(x)| = \frac{n}{2}\}$  is efficient.

Based on these observations, we can draw several conclusions concerning efficient action profiles in arbitrary network structures and their relation to the set of selfish Nash equilibria.

**Theorem 6** [EFFICIENCY OF A FULLY SATISFYING ACTION PROFILE]

- (i) For a network  $(G, \theta)$  and game  $(\delta, \lambda)$  s.t.  $\lambda > 2 \cdot |\delta - 1|$ , an action profile  $x$  is efficient if and only if it is fully satisfying. The sets of efficient and equilibrium action profiles coincide.

- (ii) For a network  $(G, \theta)$  and game  $(\delta, \lambda)$  s.t.  $\lambda < 2 \cdot (2\delta - 1)$ , if the players of the preference minority  $N^\pi$  form an independent set, then a fully satisfying action profile is not efficient.<sup>24</sup>
- (iii) For a network  $(G, \theta)$  and game  $(\delta, \lambda)$  s.t.  $\lambda > |2\delta - 1|$ , if a fully satisfying action profile is not efficient, then in any efficient action profile all frustrated players are worse off than in equilibrium.

*Proof.* (i) For given  $(G, \theta)$  and  $(\delta, \lambda)$ , denote the social welfare of an action profile  $x$  by  $U(x) = \sum_{i \in N} u_i(x)$ . Consider an arbitrary  $x$  and an arbitrary player  $i$  of degree  $d_i$ . The social welfare difference  $U(1 - \theta_i, x_{-i}) - U(\theta_i, x_{-i}) \leq |\delta - (1 - \delta)| \cdot 2d_i - \lambda \cdot d_i = (2 \cdot |2\delta - 1| - \lambda) \cdot d_i < 0$ , since  $\lambda > 2 \cdot |2\delta - 1|$  and  $d_i \geq 1 \forall i \in N$ . Since this difference is negative for an arbitrary  $x$ , it proves that a unique efficient action profile is  $x_i = \theta_i \forall i \in N$ .

(ii) Denote by  $d = \sum_{i,j \in N} \mathbf{1}_{\{\theta_i \neq \theta_j\}} / 2$  the number of links connecting players from different preference groups. Social welfare of a homogeneous action profile  $x_i = 1 - \pi \forall i \in N$  would be greater than that of a fully satisfying action profile if and only if interactional disutility from mismatching in the fully satisfying profile,  $(2\delta - 1) \cdot 2d$ , is greater than idiosyncratic disutility in the homogeneous profile,  $\lambda \cdot \sum_{i \in N^\pi} d_i$ . If players from  $N^\pi$  form an independent set, then  $\sum_{i \in N^\pi} d_i = d$ , which together with  $\lambda < 2 \cdot (2\delta - 1)$  implies that the social welfare of the homogeneous action profile  $x_i = 1 - \pi \forall i \in N$  is greater.

(iii) Let  $x^{FS}$  denote the fully satisfying action profile and  $x \neq x^{FS}$  be the efficient profile. Consider an arbitrary  $i$  of degree  $d_i$  that is frustrated in  $x$ . The utility difference  $u_i(x) - u_i(x^{FS}) \leq |2\delta - 1| \cdot d_i - \lambda \cdot d_i < 0$ , since  $\lambda > |2\delta - 1|$ , hence  $i$  is worse off in  $x$ . □

Hence, if idiosyncratic action preferences are sufficiently strong, a fully satisfying action profile is a unique efficient profile. Its efficiency is not guaranteed for weaker idiosyncratic preferences – however, it is not completely excluded. For strategic complements consider a network, in which two preference groups each form a clique and are only connected to each other by one link. If both groups are sufficiently large, a fully satisfying action profile would be the only efficient one. On the other extreme, if two preferences groups form independent sets (a bipartite network), then a fully satisfying action profile is the only efficient profile in any game of strategic substitutes.

---

<sup>24</sup>An *independent set* of nodes  $S \subset N$  in a network  $G$  is such that no two nodes in  $S$  are connected in  $G$ . For more details see Jackson (2008) or Bondy and Murty (1977).



## 2.5 Discussion

It is quite intuitive that the introduction of an additional utility component would change the set of efficient action profiles. Without idiosyncratic preferences, an efficient action profile for a game of strategic complements is the one in which all players play the same action, and for a game of strategic substitutes – the one in which the number of action mismatches is maximized. With idiosyncratic preferences, a new type of efficient action profiles appears – a fully satisfying action profile, that under specific conditions can be a unique efficient profile for games of strategic complements as well as for games of strategic substitutes, as we have seen in subsection 2.4.4.

Let us now illustrate the impact of introducing idiosyncratic action preferences on equilibrium outcomes.

It is notable that if we compare equilibrium sets for the frameworks with and without idiosyncratic preferences for the same game on the same network, there is generally no inclusion *in either direction*. On the one hand, equilibria in the framework without action preferences do not necessarily remain equilibria if we allow players to have such preferences. On the other hand, some equilibria in the framework with action preferences are never possible in the framework without.

Figure 2.8 illustrates this fact (here coloured numbers outside circles correspond to action preferences and colours of the circles – to chosen actions). Five players are connected in a network and play a game of strategic complements ( $\delta > \frac{1}{2}$ ). For the framework with idiosyncratic action preferences, we additionally assume a heterogeneous preference profile  $\theta = (0, 0, 1, 1, 0)$  and the strength of idiosyncratic preferences such that  $(\delta, \lambda) \in R_C^1(3)$ . In the framework without action preferences,  $x = (0, 0, 0, 1, 1)$  is an equilibrium (Figure 2.8(a)), while  $x = (0, 0, 1, 1, 1)$  is not (Figure 2.8(b)), as in the latter case player 3 has an incentive to deviate and follow the majority of her neighbors. However, if we allow players to have action preferences, the situation is reverse:  $x = (0, 0, 0, 1, 1)$  is not an equilibrium anymore, as player 3 has enough companions to switch to her preferred action 1 (Figure 2.8(a)), while  $x = (0, 0, 1, 1, 1)$  is an equilibrium for this very reason (Figure 2.8(b)).

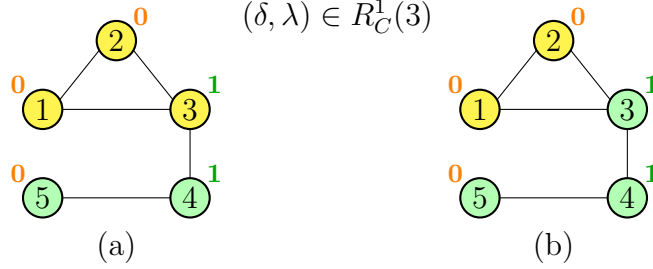


Figure 2.8: (a) is an equilibrium in the framework without idiosyncratic action preferences, but not an equilibrium in the framework with action preferences; (b) – vice versa.

Hence, we can neither claim that the introduction of idiosyncratic preferences in general extends the equilibrium set, nor that it shrinks it. What happens depends on a particular network, preference profile and strength of idiosyncratic preferences. A deeper analysis of the relationship between equilibrium sets in these two frameworks might be an interesting research question.

It is noteworthy that the above is true even in the extreme case of a homogeneous preferences profile, when all players prefer the same action. Consider a complete network with four players and  $\delta < \frac{1}{2}$ . In the framework without action preferences,  $x = (0, 0, 1, 1)$  is a unique equilibrium (up to permutation of players), since only then every player mismatches the majority of her neighbors. Now let  $\theta = (0, 0, 0, 0)$  be a preference profile. If idiosyncratic preferences are sufficiently strong,  $(\delta, \lambda) \in R_A^1(3)$ , then a unique equilibrium is  $x = (0, 0, 0, 1)$  (again, up to permutation of players). And if they are even stronger,  $(\delta, \lambda) \in R_D$ , then a unique equilibrium is  $x = (0, 0, 0, 0)$ . Intuition here is the following: when players apart from having interactional incentives are biased towards a particular action, it raises the possibility of equilibrium outcomes that are more satisfying even though less aligned with players' interactional incentives.

Let us now briefly outline the main differences in terms of results for an alternative utility function (3.1), in which idiosyncratic and interactional components are additively separable.<sup>25</sup>

Since the boundaries of the dominant action region now depend on players' degrees (see Figure 2.9), interactional incentives might not be of the same kind for all players. More precisely, in some games (blue regions on the figure) players of higher degrees seek coordination (or anti-coordination), whereas for players of lower degrees there is a dominant action – their idiosyncratic preference. In such games less connected players always satisfy their action preferences in equilibrium, which might not be the case for more connected players. Consequently, for the blue region on the right, including games of strategic complements, (at most one) homogeneous equilibrium exists if and only if all players of lower degrees prefer the same action.

<sup>25</sup>See Orlova (2019) for formal analysis.

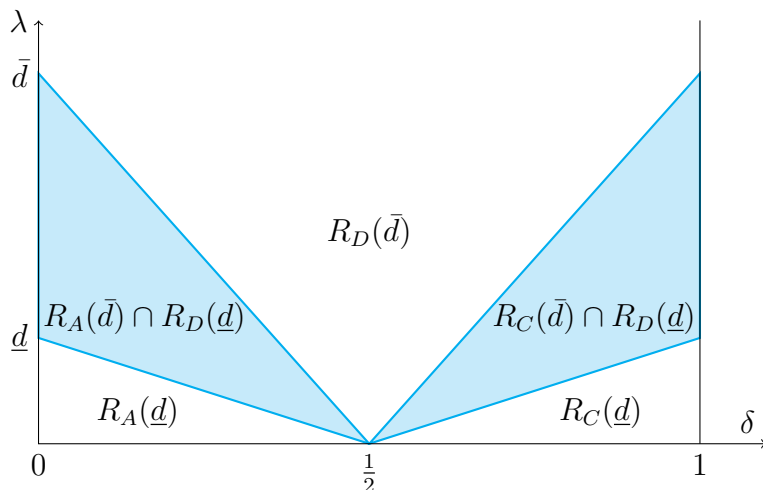


Figure 2.9: Additively separable utility. Classes of games for a network with the highest degree  $\bar{d}$  and the lowest degree  $\underline{d}$ .

Note also that in comparison to the case of interdependent utility components, the intermediate game region with a unique, fully satisfying Nash equilibrium significantly shrinks (the more the higher the degree of the most connected player). This is due to the fact that the relative weight of the idiosyncratic utility component gets lower as a player's connectivity increases, which reduces the possibility of a fully satisfying equilibrium.

## 2.6 Conclusions

Our results confirm that the introduction of idiosyncratic preferences over actions makes an important difference to equilibrium outcomes. Generically, the equilibrium set in a network where players have idiosyncratic preferences over actions is different from the analogous set in a network without such preferences, and neither is a subset of the other one. This holds even in the extreme case: the situation when players have action preferences but all prefer the same action is qualitatively different from the situation of no action preferences.

Extending the framework of Hernández et al. (2013), we characterize individual behavior and equilibrium outcomes for a large class of games. For coordination and anti-coordination games, equilibrium characterizations in cohesion terminology imply that for variation in behavior it is necessary and sufficient that the set of players can be partitioned into two sufficiently cohesive (for coordination) or sufficiently outward (for anti-coordination) subsets. What is interesting is that these subsets do not need to coincide with preferences groups, that is, a certain level of segregation (or integration) on a network is required, but not necessarily according to players' action preferences.

For dominant action games, including games of strategic complements as well as games

of strategic substitutes, a unique equilibrium exists: the one in which all players choose their preferred actions (fully satisfying equilibrium). We also derive necessary and sufficient conditions for existence of such an equilibrium in coordination and anti-coordination games. Fully satisfying equilibria in many games are unique efficient equilibria, and in yet more games they are Pareto optimal.

The main analysis is performed for interdependent relationship between idiosyncratic and interactional utility components. If these components are independent (idiosyncratic utility does not depend on the network), action preferences of less connected players are more likely to be satisfied in equilibrium, compared to more connected players. Moreover, for a range of strategic complements games only one of two homogeneous action profiles can be an equilibrium, that is, coordination becomes more selective.

## Appendix to Chapter 2

### Proof of Proposition 1

We will first prove an auxiliary lemma characterizing the best responses of players in a general form.

**Lemma 2.** *In a game  $(\delta, \lambda) \in R_C$  the best response function of a player  $i$  with preference  $\theta_i$  and  $d_i$  neighbors,  $\tau_i$  of whom play 1, is*

$$BR_i(\theta_i, d_i, \tau_i) = \begin{cases} 1, & \text{if } \tau_i > \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i) \\ 0, & \text{if } \tau_i < \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i) \\ \theta_i, & \text{if } \tau_i = \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i) \end{cases}$$

where  $\tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i) = \frac{2\delta - 1 + \lambda(1 - 2\theta_i)}{2(2\delta - 1)}d_i$ .

*Proof.* At the decision threshold the utility the player gets if she chooses action 1 should equal her utility from choosing action 0. It means that  $\delta\tau_i + (1 - \delta)(d_i - \tau_i) + d_i\lambda\theta_i = \delta(d_i - \tau_i) + (1 - \delta)\tau_i + d_i\lambda(1 - \theta_i)$ , which gives the threshold  $\tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i) = \frac{2\delta - 1 + \lambda(1 - 2\theta_i)}{2(2\delta - 1)}d_i$ .

It is straightforward to verify that in the region  $R_C$  action 0 gives higher utility whenever  $\tau_i < \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i)$ , while action 1 is preferred whenever  $\tau_i > \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i)$ . If the player is indifferent (which happens when  $\tau_i = \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i)$ ), then according to the tie-breaking rule she chooses her preferred action.  $\square$

Let us make an important observation that allows to specify further the threshold values. Since  $\tau_i$  can only be a non-negative integer, if we substitute  $\tilde{\tau}_{\delta, \lambda}^1(d_i)$  by  $\lceil \tilde{\tau}_{\delta, \lambda}^1(d_i) \rceil$  and  $\tilde{\tau}_{\delta, \lambda}^0(d_i)$  by  $\lfloor \tilde{\tau}_{\delta, \lambda}^0(d_i) \rfloor$  in the above best response function, Lemma 2 still holds.<sup>26</sup> It implies that we

<sup>26</sup>Here  $\lfloor x \rfloor$  and  $\lceil x \rceil$  denote the floor and ceiling of  $x$  respectively.

could focus solely on integer thresholds. Let us denote  $\lceil \tilde{\tau}_{\delta,\lambda}^1(d_i) \rceil$  by  $\tau_{\delta,\lambda}^1(d_i)$  and  $\lfloor \tilde{\tau}_{\delta,\lambda}^0(d_i) \rfloor$  by  $\tau_{\delta,\lambda}^0(d_i)$ . To complete the proof of the proposition we are left to verify the following:

- (i)  $\tau_{\delta,\lambda}^1(d_i) = l$  if and only if  $(\delta, \lambda) \in R_C^l(d_i)$  for  $l = 1, \dots, L$ , and
- (ii)  $\tau_{\delta,\lambda}^0(d_i) = d_i - \tau_{\delta,\lambda}^1(d_i)$ .

Let us prove the first equivalence.

Note that for every  $l = 1, \dots, L$  the condition  $\tau_{\delta,\lambda}^1(d_i) = l$  is equivalent to  $l-1 < \tilde{\tau}_{\delta,\lambda}^1(d_i) \leq l$ , which in its turn can be rewritten as a conjunction of two inequalities:

$$\begin{cases} (2\delta - 1 - \lambda)d_i > 2(2\delta - 1)(l - 1) \\ (2\delta - 1 - \lambda)d_i \leq 2(2\delta - 1)l, \end{cases}$$

or equivalently,

$$\begin{cases} \delta(2d_i - 4(l - 1)) > (1 + \lambda)d_i - 2(l - 1) & (2.3) \\ \delta(2d_i - 4l) \leq (1 + \lambda)d_i - 2l. & (2.4) \end{cases}$$

Consider any  $l = 1, \dots, L - 1$ . Since  $l \leq L - 1 = \lceil \frac{d_i}{2} \rceil - 1 < \frac{d_i}{2}$ , the above system of inequalities can be rewritten as

$$\begin{cases} \delta > \frac{(1 + \lambda)d_i - 2(l - 1)}{2(d_i - 2(l - 1))} & (2.5) \\ \delta \leq \frac{(1 + \lambda)d_i - 2l}{2(d_i - 2l)}, & (2.6) \end{cases}$$

which, provided that  $(\delta, \lambda) \in R_C$ , is precisely the condition  $(\delta, \lambda) \in R_C^l(d_i)$ . Thus, we have proved (i) for  $l = 1, \dots, L - 1$ .

Now consider  $l = L$ . Since  $\frac{d_i}{2} \leq L < \frac{d_i}{2} + 1$ , inequality (2.3) can be rewritten as (2.5) with  $l = L$ . Further, let us consider two separate cases: when  $d_i$  is even and when it is odd. If it is even, then  $L = \frac{d_i}{2}$  and (2.4) holds trivially. Therefore,  $\tau_{\delta,\lambda}^1(d_i) = L$  is equivalent to condition (2.5) with  $l = L$  and, provided that  $(\delta, \lambda) \in R_C$ , to  $(\delta, \lambda) \in R_C^L(d_i)$ . If  $d_i$  is odd, then  $\frac{d_i}{2} < L < \frac{d_i}{2} + 1$  and inequality (2.4) can be rewritten as

$$\delta \geq \frac{(1 + \lambda)d_i - 2(L - 1)}{2(d_i - 2(L - 1))}. \quad (2.7)$$

Now  $\tau_{\delta,\lambda}^1(d_i) = L$  is equivalent to the conjunction of (2.7) and (2.5) with  $l = L$ . It is straightforward to show that the right-hand-side of the former is strictly less than the right-hand-side of the latter, and thus (2.7) is redundant. Again, provided that  $(\delta, \lambda) \in R_C$ , we get that  $\tau_{\delta,\lambda}^1(d_i) = L$  is equivalent to  $(\delta, \lambda) \in R_C^L(d_i)$ , which completes the proof of (i).

Finally, (ii) follows trivially:  $\tau_{\delta,\lambda}^0(d_i) = d_i - l = d_i - \tau_{\delta,\lambda}^1(d_i)$  for every  $l = 1, \dots, L$ .  $\square$

### Proof of Corollary 1

Fix  $(\delta, \lambda) \in R_C$  and an arbitrary  $i \in N$ . We will prove the corollary by contraposition. That is, we will prove the following:  $\nexists j \in N_i(G)$  s.t.  $x_j = \theta_i \Rightarrow BR_i(\theta_i, d_i, \tau_i) = 1 - \theta_i$ .

If  $\nexists j \in N_i(G)$  s.t.  $x_j = \theta_i$ , then  $x_j = 1 - \theta_i \forall j \in N_i(G)$ . Let  $\theta_i = 0$ . Then, according to Proposition 1,  $\tau_{\delta, \lambda}^0(d_i) = d_i - l$  where  $l \in \{1, \dots, \lceil \frac{d_i}{2} \rceil\}$ , and thus  $BR_i(0, d_i, d_i) = 1$ . Let  $\theta_i = 1$ . In this case  $\tau_{\delta, \lambda}^1(d_i) = l$  where  $l \in \{1, \dots, \lceil \frac{d_i}{2} \rceil\}$ , and thus  $BR_i(1, d_i, 0) = 0$ . In either case,  $BR_i(\theta_i, d_i, \tau_i) = 1 - \theta_i$ , what was to be shown.  $\square$

### Proof of Proposition 2

Similar to the proof of Proposition 1, we will first prove an auxiliary lemma characterizing the best responses of players in a general form.

**Lemma 3.** *In a game  $(\delta, \lambda) \in R_A$  the best response function of a player  $i$  with preference  $\theta_i$  and  $d_i$  neighbors,  $\tau_i$  of whom play 1, is*

$$BR_i(\theta_i, d_i, \tau_i) = \begin{cases} 1, & \text{if } \tau_i < \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i) \\ 0, & \text{if } \tau_i > \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i) \\ \theta_i, & \text{if } \tau_i = \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i) \end{cases}$$

where  $\tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i) = \frac{2\delta - 1 + \lambda(1 - 2\theta_i)}{2(2\delta - 1)}d_i$ .

*Proof.* See the proof of Lemma 2 for derivation of the decision threshold  $\tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i)$ . It can be verified that in  $R_A$  action 1 gives higher utility whenever  $\tau_i < \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i)$  and action 0 does so whenever  $\tau_i > \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i)$ . The tie-breaking rule manages the remaining case of  $\tau_i = \tilde{\tau}_{\delta, \lambda}^{\theta_i}(d_i)$ , in which the preferred action  $\theta_i$  is chosen.  $\square$

If we substitute  $\tilde{\tau}_{\delta, \lambda}^1(d_i)$  by  $\lceil \tilde{\tau}_{\delta, \lambda}^1(d_i) \rceil$  and  $\tilde{\tau}_{\delta, \lambda}^0(d_i)$  by  $\lceil \tilde{\tau}_{\delta, \lambda}^0(d_i) \rceil$  in the above best response function, Lemma 3 still holds, so we could focus solely on integer thresholds. The rest of the proof of Proposition 2 is analogous to the above proof of Proposition 1.  $\square$

### Proof of Corollary 3

The proof uses contraposition, analogously to the proof of Corollary 1. Fix  $(\delta, \lambda) \in R_A$  and an arbitrary  $i \in N$ . If  $\nexists j \in N_i(G)$  s.t.  $x_j = 1 - \theta_i$ , then  $x_j = \theta_i \forall j \in N_i(G)$ . According to Proposition 2,  $\tau_{\delta, \lambda}^0(d_i) = l$  and  $\tau_{\delta, \lambda}^1(d_i) = d_i - l$  where  $l \in \{1, \dots, \lceil \frac{d_i}{2} \rceil\}$ . Then  $BR_i(0, d_i, 0) = 1$  and  $BR_i(1, d_i, d_i) = 0$ , that is, in either case  $BR_i(\theta_i, d_i, \tau_i) = 1 - \theta_i$ .  $\square$

### Proof of Lemma 1

- (i) Let us fix a player's degree  $d_i$ . As it follows from the definition of the partition of  $R_C$  (see subsection 2.3.1), the curve separating subregions  $R_C^m(d_i)$  and  $R_C^{m+1}(d_i)$  is

$\delta = \frac{(1+\lambda)d_i-2m}{2(d_i-2m)}$ , or equivalently,  $\lambda = \frac{(2\delta-1)(d_i-2m)}{d_i}$  on the domain  $\delta \in [\frac{1}{2}; 1]$ , which is exactly the curve  $\tilde{\lambda}(m)$  for  $\delta \geq \frac{1}{2}$ . Then  $l^* = \underset{m=1\dots L}{\operatorname{argmin}} |\lambda - \tilde{\lambda}(m)|$  corresponds to the separating curve that lies closest to a given  $\lambda$  (and separates  $R_C^{l^*}(d_i)$  and  $R_C^{l^*+1}(d_i)$ ). It can be easily seen from Figure 2.2 that if  $\lambda$  lies above this curve, it belongs to subregion  $R_C^{l^*}(d_i)$ , and if it lies below this curve, it belongs to the next subregion  $R_C^{l^*+1}(d_i)$ . Finally, let us note that  $\tilde{\lambda}(L) \leq 0$ , that is why  $\lambda$  can never lie below the curve  $\tilde{\lambda}(L)$ , and thus  $1 \leq l_{\delta,\lambda}(d_i) \leq L$ .

The proof for  $R_A$  is analogous. For a given degree  $d_i$ , the curve separating subregions  $R_A^m(d_i)$  and  $R_A^{m+1}(d_i)$  is  $\delta = \frac{(1-\lambda)d_i-2m}{2(d_i-2m)}$ , or equivalently,  $\lambda = \frac{(1-2\delta)(d_i-2m)}{d_i}$  on the domain  $\delta \in [0; \frac{1}{2}]$ , which is the curve  $\tilde{\lambda}(m)$  for  $\delta \leq \frac{1}{2}$ . Then  $l^*$  corresponds to the separating curve that lies closest to a given  $\lambda$ . If  $\lambda$  is above this curve, it belongs to subregion  $R_A^{l^*}(d_i)$ , and if it is below – to subregion  $R_A^{l^*+1}(d_i)$ . And again,  $\tilde{\lambda}(L) \leq 0$ , thus  $\lambda$  can never lie below the curve  $\tilde{\lambda}(L)$ , implying  $1 \leq l_{\delta,\lambda}(d_i) \leq L$ .

(ii) The proof follows directly from Proposition 3.

□

## Proof of Theorem 1

- (i) Take an arbitrary (connected) network  $G$  with a preference profile  $\theta$  and consider an action profile  $x = (x_1, \dots, x_n)$ . If  $x$  is homogeneous, then for every player  $i$  all her neighbors choose the same action:  $\forall i \in N \forall j \in N_i(G) x_j = x^*$  with some  $x^* \in \{0, 1\}$ . If  $x^* = 0$  then  $\tau_i = 0$ , and according to Proposition 1 player  $i$ 's best response is also 0 (since for any  $(\delta, \lambda) \in R_C$  the threshold  $\tau_{\delta,\lambda}^{\theta_i}(d_i) \geq 1$ ). If  $x^* = 1$  then  $\tau_i = d_i$ , and according to Proposition 1 player  $i$ 's best response is 1 (since the threshold  $\tau_{\delta,\lambda}^{\theta_i}(d_i) \leq d_i - 1$ ). As the above is true for all  $i \in N$ ,  $x$  is an equilibrium.
- (ii) Necessity. Fix  $(\delta, \lambda) \in R_C$ , a network  $(G, \theta)$  and let  $x = (x_1, \dots, x_n)$  be a heterogeneous equilibrium. For  $\pi = 0, 1$  set  $S^\pi := \{i \in N \mid x_i = \pi\}$ . Since  $x$  is heterogeneous, both  $S^0$  and  $S^1$  are nonempty and thus form a partition of  $N$ . We are left to prove that this partition satisfies the conditions of the theorem:

$$\forall i \in S^\pi \cap N^\pi : |N_i(G) \cap S^\pi| \geq l_{\delta,\lambda}(d_i) \text{ and}$$

$$\forall i \in S^\pi \cap N^{1-\pi} : |N_i(G) \cap S^\pi| \geq d_i - l_{\delta,\lambda}(d_i) + 1.$$

First, take a player  $i \in S^1 \cap N^\pi$ . That is,  $x_i = 1$  and  $\theta_i = \pi$  with some  $\pi \in \{0, 1\}$ . According to Proposition 1,  $BR_i(\pi, d_i, \tau_i) = 1$  iff  $\tau_i \geq \tau_{\delta,\lambda}^\pi(d_i)$  (with strict inequality if  $\pi = 0$ ). The same proposition implies that  $\tau_{\delta,\lambda}^1(d_i) = l_{\delta,\lambda}(d_i)$  and  $\tau_{\delta,\lambda}^0(d_i) = d_i - l_{\delta,\lambda}(d_i)$ . Since  $|N_i(G) \cap S^1| = \tau_i$ , it follows that  $|N_i(G) \cap S^1| \geq l_{\delta,\lambda}(d_i)$  for  $\pi = 1$  and

$|N_i(G) \cap S^1| > d_i - l_{\delta,\lambda}(d_i)$  for  $\pi = 0$ , where the last inequality can be rewritten as  $|N_i(G) \cap S^1| \geq d_i - l_{\delta,\lambda}(d_i) + 1$  (due to the fact that all terms are integers).

Second, take a player  $i \in S^0 \cap N^\pi$ . That is,  $x_i = 0$  and  $\theta_i = \pi$  with some  $\pi \in \{0, 1\}$ . Similarly, Proposition 1 implies that  $BR_i(\pi, d_i, \tau_i) = 0$  iff  $\tau_i \leq \tau_{\delta,\lambda}^\pi(d_i)$  (with strict inequality if  $\pi = 1$ ), which is equivalent to  $d_i - \tau_i \geq d_i - \tau_{\delta,\lambda}^\pi(d_i)$  (strict if  $\pi = 1$ ). Recall that  $\tau_{\delta,\lambda}^1(d_i) = l_{\delta,\lambda}(d_i)$  and  $\tau_{\delta,\lambda}^0(d_i) = d_i - l_{\delta,\lambda}(d_i)$ . Since  $|N_i(G) \cap S^0| = d_i - \tau_i$ , it follows that  $|N_i(G) \cap S^0| \geq l_{\delta,\lambda}(d_i)$  for  $\pi = 0$  and  $|N_i(G) \cap S^0| > d_i - l_{\delta,\lambda}(d_i)$  for  $\pi = 1$ , where the last inequality is equivalent to  $|N_i(G) \cap S^0| \geq d_i - l_{\delta,\lambda}(d_i) + 1$ .

Sufficiency. Fix  $(\delta, \lambda) \in R_C$  and a network  $(G, \theta)$ . Assume that there exists a partition  $\{S^0, S^1\}$  of  $N$  satisfying the conditions of the theorem and let us prove that a heterogeneous equilibrium exists. Consider an action profile  $x = (x_1, \dots, x_n)$  such that  $x_i = 0$  for  $i \in S^0$  and  $x_i = 1$  for  $i \in S^1$ . Since  $\{S^0, S^1\}$  is a partition of  $N$ , both  $S^0$  and  $S^1$  are nonempty, and thus  $x$  is a heterogeneous action profile. We are left to prove that it is an equilibrium.

Take a player  $i \in S^0$ . There are two possibilities: either  $i \in N^0$  or  $i \in N^1$ . If  $i \in S^0 \cap N^0$  then it must be that  $|N_i(G) \cap S^0| \geq l_{\delta,\lambda}(d_i)$ , which is equivalent to  $d_i - \tau_i \geq d_i - \tau_{\delta,\lambda}^0(d_i)$  (since  $\tau_{\delta,\lambda}^0(d_i) = d_i - l_{\delta,\lambda}(d_i)$ , according to Proposition 1). Then  $\tau_i \leq \tau_{\delta,\lambda}^0(d_i)$  and, again according to Proposition 1,  $BR_i(0, d_i, \tau_i) = 0$ . That is, the player  $i$  has no incentive to deviate from  $x_i = 0$ . Alternatively, if  $i \in S^0 \cap N^1$  then it must be that  $|N_i(G) \cap S^0| \geq d_i - l_{\delta,\lambda}(d_i) + 1$ , which is equivalent to  $d_i - \tau_i \geq d_i - \tau_{\delta,\lambda}^1(d_i) + 1$  (recall,  $\tau_{\delta,\lambda}^1(d_i) = l_{\delta,\lambda}(d_i)$ ), and thus  $\tau_i \leq \tau_{\delta,\lambda}^1(d_i) - 1$ . Proposition 1 implies in this case that  $BR_i(1, d_i, \tau_i) = 0$ . And again, the player  $i$  has no incentive to deviate from  $x_i = 0$ .

Now take a player  $i \in S^1$ . Either  $i \in N^0$  or  $i \in N^1$  must be true. If  $i \in S^1 \cap N^1$  then it must be that  $|N_i(G) \cap S^1| \geq l_{\delta,\lambda}(d_i)$ , which is equivalent to  $\tau_i \geq \tau_{\delta,\lambda}^1(d_i)$ . Proposition 1 implies that  $BR_i(1, d_i, \tau_i) = 1$ , and thus  $i$  has no incentive to deviate from  $x_i = 1$ . If  $i \in S^1 \cap N^0$  then it must be that  $|N_i(G) \cap S^1| \geq d_i - l_{\delta,\lambda}(d_i) + 1$ , which is equivalent to  $\tau_i \geq \tau_{\delta,\lambda}^0(d_i) + 1$ . According to Proposition 1,  $BR_i(0, d_i, \tau_i) = 1$ , implying that in this case as well  $i$  has no incentive to deviate from  $x_i = 1$ .

Since for all players their actions in  $x$  are the best responses,  $x$  is a heterogeneous equilibrium. □

## Proof of Theorem 2

Necessity. Fix  $(\delta, \lambda) \in R_C$ . Suppose that the fully satisfying action profile  $x = (\theta_1, \dots, \theta_n)$  is an equilibrium, but for some player  $i$  the condition on her neighbors does not hold:



$|N_i(G) \cap N^{\theta_i}| = |\{j \in N_i(G) : \theta_j = \theta_i\}| < l_{\delta,\lambda}(d_i)$ . Since  $x_j = \theta_j \forall j \in N$ , it implies  $|\{j \in N_i(G) : x_j = \theta_i\}| < l_{\delta,\lambda}(d_i)$ .

If  $\theta_i = 0$ , the last inequality is equivalent to  $d_i - \tau_i < l_{\delta,\lambda}(d_i)$ , or  $\tau_i > d_i - l_{\delta,\lambda}(d_i) = \tau_{\delta,\lambda}^0(d_i)$  (see Proposition 1 for the last equality), and thus  $BR_i(0, d_i, \tau_i) = 1 \neq \theta_i$  (again, from Proposition 1). If  $\theta_i = 1$  then  $\tau_i < l_{\delta,\lambda}(d_i) = \tau_{\delta,\lambda}^1(d_i)$ , implying  $BR_i(1, d_i, \tau_i) = 0 \neq \theta_i$ . In either case, the player  $i$  has an incentive to deviate from her preferred action. Hence,  $x = (\theta_1, \dots, \theta_n)$  is not an equilibrium.

Sufficiency. Fix  $(\delta, \lambda) \in R_C$  and suppose that the condition on neighbors' action preferences holds:  $|N_i(G) \cap N^{\theta_i}| = |\{j \in N_i(G) : \theta_j = \theta_i\}| \geq l_{\delta,\lambda}(d_i) \forall i \in N$ . Let us check if the fully satisfying action profile is an equilibrium. Since  $x_j = \theta_j \forall j \in N$ , the above condition implies  $|\{j \in N_i(G) : x_j = \theta_i\}| \geq l_{\delta,\lambda}(d_i) \forall i \in N$ .

Take an arbitrary  $i \in N$ . If  $\theta_i = 0$ , the above becomes  $d_i - \tau_i \geq l_{\delta,\lambda}(d_i)$ , or  $\tau_i \leq d_i - l_{\delta,\lambda}(d_i) = \tau_{\delta,\lambda}^0(d_i)$ , and Proposition 1 implies  $BR_i(0, d_i, \tau_i) = 0$ . If  $\theta_i = 1$  then  $\tau_i \geq l_{\delta,\lambda}(d_i) = \tau_{\delta,\lambda}^1(d_i)$ , and thus  $BR_i(1, d_i, \tau_i) = 1$ . In either case,  $BR_i(\theta_i, d_i, \tau_i) = \theta_i$ . Since it holds for any  $i \in N$ , the fully satisfying action profile is indeed an equilibrium.  $\square$

### Proof of Theorem 3

- (i) For an arbitrary network  $G$  with a preference profile  $\theta$  consider a homogeneous action profile  $x$ . Fix a player  $i$ . Since the action profile is homogeneous, all  $i$ 's neighbors choose the same action:  $\forall j \in N_i(G) x_j = x^*$  with some  $x^* \in \{0, 1\}$ . If  $x^* = 0$  then  $\tau_i = 0$ , and according to Proposition 2 player  $i$ 's best response is 1 (for any  $(\delta, \lambda) \in R_A$  the threshold  $\tau_{\delta,\lambda}^{\theta_i}(d_i) \geq 1$ ). If  $x^* = 1$  then  $\tau_i = d_i$ , and according to Proposition 2 player  $i$ 's best response is 0 (the threshold  $\tau_{\delta,\lambda}^{\theta_i}(d_i) \leq d_i - 1$ ). Since  $i$  has an incentive to deviate from  $x^*$ ,  $x = (x^*, \dots, x^*)$  cannot be an equilibrium action profile.

- (ii) The proof builds directly on Proposition 2 and is analogous to the proof of part (ii) of Theorem 1.  $\square$

### Proof of Theorem 4

The proof is analogous to the proof of Theorem 2 and uses the results of Proposition 2.  $\square$

### Proof of Proposition 4

Note that if player 1 is the central player,  $d_i = 1 \forall i \neq 1$  and  $l_{\delta,\lambda}(1) = 1$  for both  $R_C$  and  $R_A$ .

- (i) Fix  $(\delta, \lambda) \in R_C$ . Theorem 1 implies that homogeneous action profiles  $x_i = x_j \forall i, j \in N$  are equilibria. Furthermore, if a heterogeneous equilibrium exists, there must exist a partition  $\{S^0, S^1\}$  of  $N$  satisfying specific conditions. Suppose such a partition exists

and let  $S^\pi$  for some  $\pi \in \{0, 1\}$  include the central player. Since  $S^{1-\pi} \neq \emptyset$ ,  $\exists i \neq 1$  s.t.  $i \in S^{1-\pi}$ . Then  $|N_i(g) \cap S^{1-\pi}| = 0$ , which contradicts the conditions of Theorem 1 (ii). Thus, no heterogeneous equilibria exist.

- (ii) Fix  $(\delta, \lambda) \in R_A$ . Theorem 3 implies that a (heterogeneous) equilibrium exists iff there exists a partition  $\{S^0, S^1\}$  of  $N$  satisfying specific conditions. Suppose such a partition exists and let  $S^\pi$  for some  $\pi \in \{0, 1\}$  include the central player. If  $i \in S^{1-\pi}$ ,  $|N_i(g) \cap S^\pi| = 1 \geq 1$  and thus satisfies conditions of Theorem 3 (ii). If  $i \in S^\pi$  and  $i \neq 1$ ,  $|N_i(G) \cap S^{1-\pi}| = 0$ , which does not satisfy the conditions. Thus,  $S^\pi = \{1\}$  and  $S^{1-\pi} = N \setminus \{1\}$ . Finally, for the central player  $|N_i(G) \cap S^{1-\pi}| = n - 1$ , which satisfies both conditions of Theorem 3 (ii) (that correspond to the cases  $\theta_1 = \pi$  and  $\theta_1 = 1 - \pi$ ).

□

### Proof of Proposition 5

Given  $\theta$ , let  $m$  be the number of peripheral players with the same action preference as the central player:  $m = |\{i \in N : i \neq 1, \theta_i = \theta_1\}|$ . For every action profile  $x$ , let us define the following two numbers:  $\tau_s = |\{i \in N : i \neq 1, x_i = \theta_1 = \theta_i\}|$  and  $\tau_f = |\{i \in N : i \neq 1, x_i = \theta_1 \neq \theta_i\}|$ , representing peripheral players who choose the action preferred by the central player and are satisfied or frustrated respectively. Obviously,  $\tau_s \in \{0, \dots, m\}$  and  $\tau_f \in \{0, \dots, n - 1 - m\}$ .

The social welfare of an action profile  $x$  is determined by these two numbers and the action of the central player:

$$U(x_1, \tau_s, \tau_f) = \begin{cases} 2(\delta(\tau_s + \tau_f) + (1 - \delta)(n - 1 - \tau_s - \tau_f)) + \lambda(\tau_s + n - 1 - m - \tau_f + n - 1) & \text{if } x_1 = \theta_1, \\ 2(\delta(n - 1 - \tau_s - \tau_f) + (1 - \delta)(\tau_s + \tau_f)) + \lambda(\tau_s + n - 1 - m - \tau_f) & \text{if } x_1 \neq \theta_1, \end{cases}$$

or, equivalently,

$$U(x_1, \tau_s, \tau_f) = \begin{cases} [2(2\delta - 1) + \lambda] \cdot \tau_s + [2(2\delta - 1) - \lambda] \cdot \tau_f + 2(1 - \delta + \lambda)(n - 1) - \lambda m & \text{if } x_1 = \theta_1, \\ [2(1 - 2\delta) + \lambda] \cdot \tau_s + [2(1 - 2\delta) - \lambda] \cdot \tau_f + (2\delta + \lambda)(n - 1) - \lambda m & \text{if } x_1 \neq \theta_1. \end{cases}$$

Maximizing this function with respect to  $\tau_s$ ,  $\tau_f$  and  $x_1$  on the corresponding domain gives

for  $\lambda < 2(2\delta - 1)$ :  $\tau_s^* = m, \tau_f^* = n - 1 - m, x_1^* = \theta_1$  (and, if  $m = 0$ :  $\tau_s^* = 0, \tau_f^* = 0, x_1^* \neq \theta_1$ ),

for  $\lambda < 2(1 - 2\delta)$ :  $\tau_s^* = 0, \tau_f^* = 0, x_1^* = \theta_1$  (and, if  $m = n - 1$ :  $\tau_s^* = n - 1, \tau_f^* = 0, x_1^* \neq \theta_1$ ),

and for  $\lambda > 2|2\delta - 1|$ :  $\tau_s^* = m, \tau_f^* = 0, x_1^* = \theta_1$ . □

### Proof of Proposition 6

- (i) Fix  $(\delta, \lambda) \in R_C$ . Theorem 1 implies that homogeneous action profiles  $x_i = x_j \forall i, j \in N$  are equilibria. Furthermore, other (heterogeneous) equilibria exist iff there exists a

partition  $\{S^0, S^1\}$  of  $N$ , which for each  $\pi \in \{0, 1\}$  satisfies at least one of the following conditions:  $|S^\pi| \geq l + 1$  or  $|S^\pi| \geq n - l + 1$ . Note that the second condition is at least as strong as the first one:  $l \leq \lceil \frac{n-1}{2} \rceil$  implies  $l + 1 \leq \lceil \frac{n+1}{2} \rceil$  and  $n - l + 1 \geq \lceil \frac{n+1}{2} \rceil$ . There are three possibilities. First, when only the first condition holds for  $\pi = 0, 1$ , subsets  $S^0 \cap N^1$  and  $S^1 \cap N^0$  are empty, that is,  $S^0 = N^0$  and  $S^1 = N^1$ . Thus, the fully satisfying action profile  $x_i = \theta_i \forall i \in N$  is an equilibrium when  $n^\pi \geq l + 1 \forall \pi \in \{0, 1\}$ . Second, when only the first condition holds for some  $\pi \in \{0, 1\}$  but for  $1 - \pi$  the second condition also holds,  $|S^0| + |S^1| \geq l + 1 + n - l + 1 = n + 2$ , which contradicts  $|S^0| + |S^1| = n$ . Thus, no other heterogeneous equilibria exist in this case. Finally, when the second (stronger) condition holds for  $\pi = 0, 1$ , the sum  $|S^0| + |S^1|$  is even larger than in the previous case, which again contradicts  $|S^0| + |S^1| = n$ . It proves that no other equilibria are possible.

- (ii) Fix  $(\delta, \lambda) \in R_A$ . Theorem 3 implies that a (heterogeneous) equilibrium exists iff there exists a partition  $\{S^0, S^1\}$  of  $N$ , which for each  $\pi \in \{0, 1\}$  satisfies at least one of the following conditions:  $|S^\pi| \geq l$  or  $|S^\pi| \geq n - l$ . Note that the second condition is at least as strong as the first one:  $l \leq \lceil \frac{n-1}{2} \rceil$  implies  $n - l \geq \lceil \frac{n-1}{2} \rceil$ . There are three possibilities. First, when only the first condition holds for  $\pi = 0, 1$ , subsets  $S^0 \cap N^1$  and  $S^1 \cap N^0$  are empty, that is,  $S^0 = N^0$  and  $S^1 = N^1$ . Thus, the fully satisfying action profile  $x_i = \theta_i \forall i \in N$  is an equilibrium when  $n^\pi \geq l \forall \pi \in \{0, 1\}$ . Second, when only the first condition holds for some  $\pi \in \{0, 1\}$  but for  $1 - \pi$  the second condition also holds ( $|S^\pi| \geq l$  and  $|S^{1-\pi}| \geq n - l$ ), it must be that  $|S^\pi| = l$ , otherwise it would contradict  $|S^0| + |S^1| = n$ . Moreover,  $S^{1-\pi} \cap N^\pi = \emptyset$  implies  $N^\pi \subset S^\pi$ . Finally, when the second condition holds for  $\pi = 0, 1$ , it must be that the sum  $|S^0| + |S^1| \geq 2(n - l)$ , which implies  $l \geq \frac{n}{2}$ , and together with  $l \leq \lceil \frac{n-1}{2} \rceil$  implies  $l = \frac{n}{2}$ . Hence,  $|S^0| = |S^1| = \frac{n}{2}$  in this case.

□

### Proof of Proposition 7

For an action profile  $x$ , define  $S^\pi(x) = \{i \in N : x_i = \pi\}$  for  $\pi = 0, 1$ . Note that  $N$  can be decomposed into four subsets:  $S^1(x) \cap N^1$ ,  $S^1(x) \cap N^0$ ,  $S^0(x) \cap N^1$  and  $S^0(x) \cap N^0$ . Denote the number of players in the first two subsets by  $\tau_s$  and  $\tau_f$  respectively, then the second two subsets contain  $n^1 - \tau_s$  and  $n^0 - \tau_f$  players. Respective players' utilities in these subsets equal  $\delta(\tau_s + \tau_f - 1) + (1 - \delta)(n - \tau_s - \tau_f) + \lambda(n - 1)$ ,  $\delta(\tau_s + \tau_f - 1) + (1 - \delta)(n - \tau_s - \tau_f)$ ,  $\delta(n - \tau_s - \tau_f - 1) + (1 - \delta)(\tau_s + \tau_f)$  and  $\delta(n - \tau_s - \tau_f - 1) + (1 - \delta)(\tau_s + \tau_f) + \lambda(n - 1)$ . Adding up players' utilities multiplied by respective subsets' cardinalities gives social welfare of an action profile  $x$ , determined by  $\tau_s$  and  $\tau_f$ :

$$U(\tau_s, \tau_f) = 2(2\delta - 1)(\tau_s + \tau_f)^2 - 2n(2\delta - 1)(\tau_s + \tau_f) + \lambda(n - 1)(\tau_s - \tau_f) + (n - 1)(n\delta + n^0\lambda).$$

Maximization of this function with respect to  $\tau_s$  and  $\tau_f$  on the corresponding domain,  $\tau_s \in \{0, \dots, n^1\}$  and  $\tau_f \in \{0, \dots, n^0\}$ , can be simplified by means of a variable change:  $x = \tau_s + \tau_f$ ,  $y = \tau_s - \tau_f$ . After subtracting the constant, the simplified objective function becomes  $U(x, y) = 2(2\delta - 1)x^2 - 2n(2\delta - 1)x + \lambda(n - 1)y$ , where  $0 \leq x + y \leq 2n^1$  and  $0 \leq x - y \leq 2n^0$ . Note that optimal  $y^*(x) = \begin{cases} x & \text{if } x \in \{0, \dots, n^1\} \\ -x + 2n^1 & \text{if } x \in \{n^1, \dots, n\} \end{cases}$ , so that we can perform maximization piecewise. Without loss of generality, assume  $n^1 \leq n^0$ . We have to consider the cases  $\delta \geq \frac{1}{2}$  and  $\delta \leq \frac{1}{2}$  separately.

Let  $\delta \geq \frac{1}{2}$ . It is easy to check that  $x^* = y^* = 0$  (implying  $\tau_s^* = \tau_f^* = 0$ ) for  $\lambda \leq 2(2\delta - 1)\frac{n^0}{n-1}$ , and  $x^* = y^* = n^1$  (implying  $\tau_s^* = n^1, \tau_f^* = 0$ ) for  $\lambda \geq 2(2\delta - 1)\frac{n^0}{n-1}$ . Let now  $\delta \leq \frac{1}{2}$ . Then  $x^* = y^* = \left\lfloor \frac{n}{2} - \frac{\lambda(n-1)}{4(1-2\delta)} \right\rfloor$  (and thus  $\tau_s^* = \left\lfloor \frac{n}{2} - \frac{\lambda(n-1)}{4(1-2\delta)} \right\rfloor, \tau_f^* = 0$ ) for  $\lambda \leq 2(1 - 2\delta)\frac{n-2n^1}{n-1}$ , and  $x^* = y^* = n^1$  (and thus  $\tau_s^* = n^1, \tau_f^* = 0$ ) for  $\lambda \geq 2(1 - 2\delta)\frac{n-2n^1}{n-1}$ .  $\square$

# Chapter 3

## Network Games with Heterogeneous Players

**Abstract:** In this paper we consider network games in which players simultaneously form partnerships and choose actions. Players are heterogeneous with respect to their action preferences. We characterize pairwise Nash equilibria for a large class of games, including coordination and anti-coordination games, varying the strength of action preferences and the size of the linking cost. We find that, despite the symmetry and simplicity of the setting, quite irregular network structures can arise in equilibrium, implying that heterogeneity in players' action preferences may already explain a large part of observed irregularity in endogenously formed networks.

**JEL codes:** C62, C72, D85.

**Keywords:** *network games; strategic network formation; preference heterogeneity; efficiency.*

### 3.1 Introduction

In social contexts, an individual's choice is often strongly influenced by choices of other related to her individuals. This social influence is frequently modeled as a non-cooperative game played on a fixed network, where each individual plays a common bilateral game with each of her network partners and obtains the sum of these bilateral games' payoffs. Games on networks were first systematically introduced in Galeotti et al. (2010) and have been actively studied since then (see a recent overview of Bramoullé and Kranton (2016)). However, quite often individuals also have considerable control over whom they interact with. The first models of strategic network formation date back to Myerson (1977) and are more recently

surveyed, for instance, in Goyal (2016) and Mauleon and Vannetelbosch (2016). These two strands of research – network formation and games on networks – have been subsequently combined in models that consider games on endogenous networks. A good overview of the literature that studies the interplay between individual behavior and the formation of an interactional structure is Vega-Redondo (2016). Our paper contributes to this literature.

We investigate one particular aspect – the impact of *ex ante* heterogeneity between players. In particular, we allow players differ in their action preferences. On a fixed network, this would often create a conflict between a player’s idiosyncratic action preference and her interactional incentives dictated by action choices of her network partners. If the network is endogenous, however, a player might just choose not to interact with those whose actions do not correspond to her own preferred action.<sup>1</sup> Ellwardt et al. (2016) and Goyal et al. (2021) show experimentally that this is a typical outcome in a two-stage coordination game under complete information, when individuals form their partnerships prior to choosing their actions. Goyal et al. (2021) also check the robustness of these results to non-zero values of the linking cost. Our aim is to derive equilibrium characterizations analytically and for a considerably larger class of games, varying also the strength of individuals’ action preferences and the size of the linking cost. For this purpose, we extend the theoretical framework of Orlova (2019) from games with heterogeneous players on a fixed network to games on an endogenous network.

The setting is the following. We consider network games in which players with heterogeneous preferences over actions simultaneously form a network and choose their actions. The action choice is binary and hence there are two types of players. If a player chooses her preferred action, she gets a higher payoff in every bilateral game she plays. Link formation is two-sided, that is, links are formed between those players who have made mutual link proposals. Both link proposals and link maintenance are costly. The same bilateral game is played between all pairs of players who decided to be linked; it can be either a coordination game, an anti-coordination game, or a dominant action game (if individuals’ action preferences are very strong). We consider a complete information setting and use a static solution concept – pairwise Nash equilibrium. The implications of alternative equilibrium concepts are also discussed.

We find that, despite relative simplicity and symmetry of the setting (ex ante there are only two types of players that differ in their action preferences), quite irregular network structures are possible in equilibrium (see Table 3.1). These are partially connected networks with heterogeneous action profiles such that only a part of players choose their preferred actions. Such irregular equilibrium structures might exist both for coordination and for anti-

---

<sup>1</sup>This concerns two-sided link formation models, in which every link requires an agreement of both involved partners but can be severed unilaterally.

coordination games, and it can be shown that their existence is robust to some equilibrium refinements – for instance, a bilateral Nash equilibrium.

The rest of the paper is organized as follows. Section 2 describes the model and provides all necessary definitions. Section 3 presents the results, proposes a classification of equilibria with respect to the action profile and the network structure and illustrates them with examples. Section 4 highlights the impact of heterogeneity on equilibrium outcomes, discusses alternative equilibrium concepts and describes planned follow-up research on efficiency of the derived equilibria. Appendix contains the proofs of all the results.

## 3.2 The model

### 3.2.1 The game

Let  $N = \{1, \dots, n\}$  be the set of players and  $\theta = (\theta_1, \dots, \theta_n)$  be the *preference profile* of players, where  $\theta_i \in \{0, 1\} \forall i \in N$ . For  $\pi \in \{0, 1\}$  we call  $N^\pi = \{i \in N \mid \theta_i = \pi\}$  a *preference group* with action preference  $\pi$  and denote its cardinality by  $n^\pi$ . We assume that  $n^\pi \geq 2 \forall \pi \in \{0, 1\}$ , that is, we consider games with heterogeneous preference profiles.

Each player simultaneously chooses a (pure) strategy  $s_i = (x_i, p_i) \in S_i = \{0, 1\}^n$  consisting of an action  $x_i \in \{0, 1\}$  and a vector of link proposals to other players  $p_i = (p_{i1}, \dots, p_{i,i-1}, p_{i,i+1}, \dots, p_{in}) \in \{0, 1\}^{n-1}$ . Any strategy profile  $s \in S = S_1 \times \dots \times S_n$  induces a directed graph of proposals  $P$ , which can be represented by an adjacency matrix:  $P_{ij} = p_{ij} \forall i \neq j$  and  $P_{ii} = 0 \forall i \in N$ .<sup>2</sup> The links are formed between those players who made mutual proposals, inducing an undirected graph (network)  $G$  with  $G_{ij} = P_{ij} \cdot P_{ji} \forall i, j \in N$ .<sup>3</sup>

We denote by  $\bar{S}$  the subset of strategy profiles that do not contain unreciprocated proposals:  $\bar{S} = \{s \in S \mid p_{ij} = p_{ji} \forall i, j \in N\}$ . In what follows  $s_{-i}$  designates the strategy vector of all players except for  $i$  and  $s_{-i-j}$  the strategy vector of all players except for  $i$  and  $j$ . For a given  $s_{-i}$ , we denote by  $\bar{S}_i(s_{-i})$  all  $i$ 's strategies that do not contain  $i$ 's unreciprocated proposals:  $\bar{S}_i(s_{-i}) = \{s_i \in S_i \mid p_{ij} = 1 \Rightarrow p_{ji} = 1 \forall j \in N\}$ . Obviously, for  $s \in \bar{S}$  it holds that  $s_i \in \bar{S}_i(s_{-i}) \forall i \in N$ .

The payoff for a player  $i$  with action preference  $\theta_i$  is

$$u_i(s) = \sum_{j \in N} p_{ij} p_{ji} (\delta \cdot \mathbf{1}_{\{x_i = x_j\}} + (1 - \delta) \cdot \mathbf{1}_{\{x_i \neq x_j\}} + \lambda \cdot \mathbf{1}_{\{x_i = \theta_i\}} - (c - \varepsilon)) - \varepsilon \cdot \sum_{j \in N} p_{ij},$$

where  $\delta \in [0; 1]$ ,  $\lambda \in [0; +\infty)$  and  $c > \varepsilon > 0$ .

<sup>2</sup>By convention, players do not make link proposals to themselves. Note that link proposals  $p_{ij}$  are defined only for such  $i, j \in N$  that  $i \neq j$ .

<sup>3</sup>The terms *network* and (undirected) *graph* are used interchangeably in this paper.

Hence, a player enjoys network benefits from her connections in the induced network  $G$ , while she has to pay a positive cost  $\varepsilon$  for each link proposal and a positive link maintenance cost  $c - \varepsilon$  for each link. Note that  $i$ 's network benefits consist of two parts: interactional benefits, that depend on the actions chosen by  $i$ 's network neighbors (parameter  $\delta$  determines relative advantage of matching versus mismatching actions), and idiosyncratic benefits, that arise if  $i$  chooses her preferred action  $\theta_i$  (parameter  $\lambda$  determines the strength of action preferences). We analyze a class of games  $\Gamma = \{\Gamma_{\delta,\lambda} \mid 0 \leq \delta \leq 1, \lambda \geq 0\}$ , where every specific game is determined by two parameters (with a slight abuse of terminology, we will sometimes refer to a pair  $(\delta, \lambda)$  as "a game", implying the corresponding  $\Gamma_{\delta,\lambda}$ ). Depending on the relative values of these parameters, any game  $\Gamma_{\delta,\lambda}$  can be classified into one of the following subclasses: coordination games, anti-coordination games or dominant action games (see Figure 3.1).

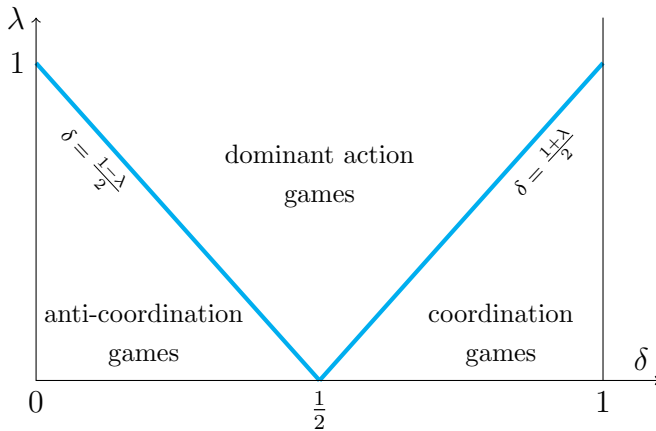


Figure 3.1: Parameter regions, representing three subclasses of games.

Note that if a player's strategy does not contain unreciprocated proposals, i.e.  $s_i \in \bar{S}_i(s_{-i})$ , then her payoff function can be simplified:

$$\begin{aligned}
 u_i(s) &= \sum_{j \in N} p_{ij} p_{ji} (\delta \cdot \mathbb{1}_{\{x_i=x_j\}} + (1 - \delta) \cdot \mathbb{1}_{\{x_i \neq x_j\}} + \lambda \cdot \mathbb{1}_{\{x_i=\theta_i\}} - c) \\
 &= \sum_{j \in N} p_{ij} p_{ji} u_{ij}(x_i, x_j),
 \end{aligned} \tag{3.1}$$

where  $u_{ij}(x_i, x_j)$  denotes  $i$ 's payoff component due to her link with  $j$ .<sup>4</sup> The total linking cost  $c$ , that a player pays for each of her links, combines the cost of link proposal and the cost of link maintenance.

<sup>4</sup>Note that for any pair of connected players  $u_{ij}(x_i, x_j)$  has only four possible values:  $\delta - c$ ,  $1 - \delta - c$ ,  $\delta + \lambda - c$  or  $1 - \delta + \lambda - c$ .



We consider a complete information setting, that is, the players' preference profile and their payoff functions are common knowledge prior to the game. Players choose their strategies simultaneously, aiming to maximize their respective payoffs.

### 3.2.2 Equilibrium concept and some graph theory notions

Consider a game  $\Gamma_{\delta,\lambda}$  and fix some linking cost  $c > 0$ . A strategy profile  $s$  is a Nash equilibrium (NE) of the game if and only if  $\forall i \in N \forall s'_i \in S_i u_i(s'_i, s_{-i}) \leq u_i(s)$ .<sup>5</sup> In the spirit of the networks literature, we refine the set of Nash equilibria by introducing pairwise Nash equilibria. We allow pairs of unlinked players to deviate cooperatively by creating a mutual link with a possibility to simultaneously adjust their action choices. Formally, for a strategy profile  $s \in S$  and a pair of players  $i, j \in N$  s.t.  $p_{ij}p_{ji} = 0$ , a *pairwise deviation*  $((x'_i, p'_i), (x'_j, p'_j))$  is such a deviation that  $p'_{ij}p'_{ji} = 1$  and  $p'_{kl} = p_{kl} \forall k \in \{i, j\} \forall l \notin \{i, j\}$ . A pairwise Nash equilibrium is a Nash equilibrium proof against such pairwise deviations.<sup>6</sup>

**Definition 1.** A strategy profile  $s = (x, p)$  is a *pairwise Nash equilibrium (PNE)* of the above game if it is a Nash equilibrium and for any pair  $i, j \in N$  s.t.  $p_{ij}p_{ji} = 0$  and any  $x'_i, x'_j \in \{0, 1\}$ ,

$$u_i((x'_i, p'_i), (x'_j, p'_j), s_{-i-j}) > u_i(s) \Rightarrow u_j((x'_i, p'_i), (x'_j, p'_j), s_{-i-j}) < u_j(s),$$

where  $p'_{ij}p'_{ji} = 1$  and  $p'_{kl} = p_{kl} \forall k \in \{i, j\} \forall l \notin \{i, j\}$ .

For a given game we denote by  $S^{PNE} \subseteq S^{NE}$  the sets of pairwise Nash equilibria and Nash equilibria respectively. In the following section, we analyze pairwise Nash equilibria for different games  $\Gamma_{\delta,\lambda} \in \Gamma$  and different sizes of the linking cost  $c$ .

Before we move to equilibrium characterizations, let us remind several definitions from the graph theory that will appear useful in our analysis.<sup>7</sup>

A graph  $G$  in which each pair of distinct nodes is linked,  $G_{ij} = 1 \forall i \neq j$ , is called a *complete* graph. An *empty* graph, on the other hand, is one with no links:  $G_{ij} = 0 \forall i, j \in N$ . A *bipartite* graph is one that admits a partition of its set of nodes  $N$  into two subsets  $N'$  and  $N''$  in such a way that every link of  $G$  connects a node of  $N'$  and a node of  $N''$ :  $G_{ij} = 1 \Rightarrow (i \in N' \wedge j \in N'') \vee (i \in N'' \wedge j \in N')$ . In a *complete bipartite* graph every node of  $N'$  is linked to every node of  $N''$ :  $G_{ij} = 1 \Leftrightarrow (i \in N' \wedge j \in N'') \vee (i \in N'' \wedge j \in N')$ .

<sup>5</sup>Since only pure strategies are admissible, all equilibria in this paper are pure strategy equilibria.

<sup>6</sup>The same definition appears in Hiller (2017).

<sup>7</sup>The following definitions are based on Bondy and Murty (1977), Diestel (2017) and Benjamin et al. (2015). The terms *vertex* and *edge* are substituted by more common in the networks literature terms *node* and *link* respectively.

A graph  $G'$  is called a subgraph of a graph  $G$  if every node and link of  $G'$  is a node and link, respectively, of  $G$ . A graph  $G$  is a subgraph of itself; all other subgraphs are proper subgraphs of  $G$ . If a proper subgraph  $G' \subset G$  is complete, it is called a *clique*.<sup>8</sup> Let  $G'$  and  $G''$  be proper subgraphs of  $G$  with corresponding sets of nodes  $N'$  and  $N''$ . We say that  $G'$  and  $G''$  are *disjoint* if they have no nodes in common:  $N' \cap N'' = \emptyset$ . We say that disjoint  $G', G'' \subset G$  are *connected*, if  $\exists i \in N' \exists j \in N''$  s.t.  $G_{ij} = 1$ , otherwise they are *disconnected*.

Finally, let  $G'$  and  $G''$  be two graphs. A *union* of  $G'$  and  $G''$  is a graph with the set of nodes  $N = N' \cup N''$  and links such that  $G_{ij} = 1 \Leftrightarrow G'_{ij} = 1 \vee G''_{ij} = 1 \forall i, j \in N$ .

### 3.3 Equilibrium analysis

#### 3.3.1 Preliminaries

This subsection establishes important relations between certain sets of strategy profiles and then formulates necessary and sufficient conditions for a strategy profile to be a pairwise Nash equilibrium.

First, fix any game  $\Gamma_{\delta, \lambda} \in \Gamma$ . Without loss of generality, let us make a technical assumption about admissible values of the linking cost.

*Assumption 1.* Given a game  $\Gamma_{\delta, \lambda}$ , a linking cost  $c$  can take any values in  $C_{\delta, \lambda} := \mathbb{R}_{++} \setminus \{\delta, 1 - \delta, \delta + \lambda, 1 - \delta + \lambda\}$ .

That is, a linking cost  $c$  can take any positive real values, except for four specific ones.<sup>9</sup> Taking into account this assumption, fix any linking cost  $c$ . The first lemma relates Nash equilibria of a game to the set  $\bar{S}$  of strategy profiles without unreciprocated proposals.

**Lemma 1.**  $S^{NE} \subseteq \bar{S}$ .

In other words, a Nash equilibrium cannot contain unreciprocated proposals. This directly follows from the fact that every link proposal carries a strictly positive cost. Formal proofs of this and the following lemmas are moved to the appendix.

Next, we notice that not only Nash equilibria do not contain unreciprocated proposals, but also profitable unilateral strategy deviations cannot contain unreciprocated proposals of the deviating player. This leads to an alternative characterization of the Nash equilibrium set.

---

<sup>8</sup>Note that this definition is different from another common one that appears, for instance, in Jackson (2008) and defines a clique as a *maximal* completely connected subgraph of  $G$ .

<sup>9</sup>This assumption guarantees that players are never indifferent to any of their links, that is,  $u_{ij}(x_i, x_j) \neq 0 \forall i, j \in N \forall x_i, x_j \in \{0, 1\}$  (see footnote 4). If this assumption does not hold, more equilibria are possible, but none of them is robust to small changes in parameter values.

**Lemma 2.**  $s \in S^{NE}$  if and only if  $\forall i \in N \forall s'_i \in \bar{S}_i(s_{-i}) u_i(s'_i, s_{-i}) \leq u_i(s)$ .

Compared to the original definition, this one narrows down the set of relevant deviations and thus simplifies the search of equilibria.

Consider now the following set:  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow (u_{ij}(x_i, x_j) > 0 \wedge u_{ji}(x_j, x_i) > 0)\}$ . It is the subset of strategy profiles without unreciprocated proposals in which two players are linked if and only if they both benefit from the link. Lemma 3 states that all pairwise Nash equilibria must be in this set.

**Lemma 3.**  $S^{PNE} \subseteq \bar{S}^+$ .

This allows us to consider  $\bar{S}^+$  as a pool of candidate equilibrium profiles. Note, however, that  $S^{NE} \subseteq \bar{S}^+$  does not have to hold. The next lemma establishes necessary and sufficient conditions for  $s \in \bar{S}^+$  to be a Nash equilibrium. In what follows we denote by  $\tilde{s}_i = (\tilde{x}_i, \tilde{p}_i)$  a particular unilateral deviation of player  $i$  from her strategy in the strategy profile  $s = (x, p)$ :

$\tilde{x}_i \neq x_i$  and  $\tilde{p}_{ij} = \begin{cases} 0 & \text{if } u_{ij}(\tilde{x}_i, x_j) < 0 \\ p_{ij} & \text{otherwise} \end{cases}$ . In this deviation, a player  $i$  changes her action

and withdraws all her proposals for those links that are no longer profitable for  $i$ .

**Lemma 4.** Let  $s \in \bar{S}^+$ . Then  $s \in S^{NE}$  if and only if  $\forall i \in N u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$ .

Hence, for every player  $i$ ,  $\tilde{s}_i$  is the most successful of all possible unilateral deviations. If  $s \in \bar{S}^+$  is proof against such deviations, it is also proof against all other unilateral deviations.

Building upon this result, the next lemma provides necessary and sufficient conditions for  $s \in \bar{S}^+$  to be a pairwise Nash equilibrium. These conditions include proofness against three additional, pairwise deviations.

**Lemma 5.**  $s \in S^{PNE}$  if and only if  $s \in \bar{S}^+$  and the following conditions hold for all  $i \in N$  and all  $j \in N$  s.t.  $p_{ij}p_{ji} = 0$ :

$$(1) u_i(\tilde{s}_i, s_{-i}) \leq u_i(s),$$

$$(2) u_i((x'_i, p'_i), (x'_j, p'_j), s_{-i-j}) > u_i(s) \Rightarrow u_j((x'_i, p'_i), (x'_j, p'_j), s_{-i-j}) < u_j(s),$$

where either  $x'_i \neq x_i$  or  $x'_j \neq x_j$ ,  $p'_{ij}p'_{ji} = 1$  and  $p'_{kl} = p_{kl} \forall k \in \{i, j\} \forall l \notin \{i, j\}$ .

Hence, we derived necessary and sufficient conditions for a strategy profile to be a PNE: a strategy profile must contain only reciprocated proposals, must induce a link if and only if both linked players benefit from it, and must be proof against four specific (one unilateral and three pairwise) strategy deviations. Note that since not only original strategies  $s_i$  but also all relevant strategy deviations do not contain  $i$ 's unreciprocated proposals, we can use the utility function (3.1).

Finally, let us further simplify necessary and sufficient conditions for PNE for a specific (actually, very broad) range of parameter values. Denote by  $C_{\delta,\lambda}^h$  the subset of  $C_{\delta,\lambda}$  that corresponds to high values of the linking cost (see Figure 3.2 in the following subsection):  $C_{\delta,\lambda}^h = \{c \in C_{\delta,\lambda} \mid \max\{\delta, 1 - \delta + \lambda\} < c < \delta + \lambda \vee \max\{1 - \delta, \delta + \lambda\} < c < 1 - \delta + \lambda\}$ . The final lemma characterizes pairwise Nash equilibria when the linking cost is not high.

**Lemma 6.** *Let  $c \notin C_{\delta,\lambda}^h$ . Then  $s \in S^{PNE}$  if and only if  $s \in \bar{S}^+$  and the following conditions hold for all  $i \in N$  and for all  $j \in N$  s.t.  $p_{ij}p_{ji} = 0$ :*

$$(1) \quad u_i(\tilde{s}_i, s_{-i}) \leq u_i(s),$$

$$(2) \quad u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) < u_i(s),$$

where  $\hat{x}_i \neq x_i$ ,  $\hat{p}_{ij}\hat{p}_{ji} = 1$  and  $\hat{p}_{kl} = p_{kl} \forall k \in \{i, j\} \forall l \notin \{i, j\}$ .<sup>10</sup>

This characterization differs from the one in Lemma 5 by requiring to consider yet fewer pairwise deviations (two for each unlinked pair of players). In these deviations only one of the players changes her action, and this same player must bear utility loss from such a deviation. Lemmas 5 and 6 will be used extensively to prove the results of the next subsection.

### 3.3.2 Classes of equilibria

The following figure depicts ten regions of parameter values – a game  $(\delta, \lambda)$  and a linking cost  $c$  – that correspond to qualitatively different equilibrium sets. In each region only specific classes of equilibria are possible.

---

<sup>10</sup>It can be shown that if  $c \in C_{\delta,\lambda}^h$  then these conditions are necessary but not sufficient for  $s \in S^{PNE}$ .

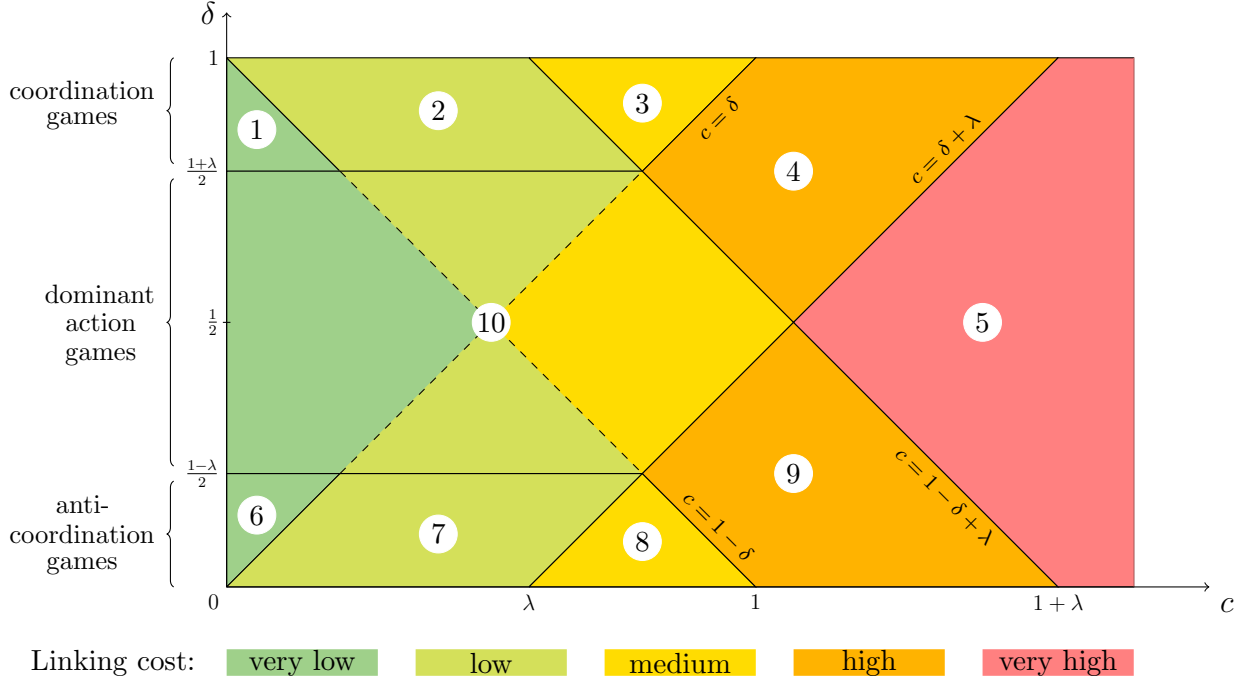


Figure 3.2: Regions of the linking cost values that correspond to different equilibrium sets. Each horizontal section defines a game  $(\delta, \lambda)$ .

With respect to the equilibrium action profile, we differentiate between PNE with a homogeneous action profile ( $x_i = x_j \forall i, j \in N$ ), PNE with a so-called fully satisfying action profile ( $x_i = \theta_i \forall i \in N$ ) and the remaining ones – PNE with a heterogeneous not fully satisfying action profile.

With respect to the equilibrium network structure, all PNE appear to fall into one of the following six classes: such that induce an empty network ( $G_{ij} = 0 \forall i, j \in N$ ), a complete network ( $G_{ij} = 1 \forall i, j \in N$ ), a network consisting of two disconnected disjoint cliques (more specifically,  $G_{ij} = 1 \Leftrightarrow x_i = x_j$ ), a network consisting of two connected disjoint cliques ( $G_{ij} = 1 \Leftrightarrow x_i = x_j \vee \theta_i = x_i \neq x_j = \theta_j$ ), a complete bipartite network ( $G_{ij} = 1 \Leftrightarrow x_i \neq x_j$ ) or a union of a complete bipartite network and a clique ( $G_{ij} = 1 \Leftrightarrow x_i \neq x_j \vee \theta_i = x_i = x_j = \theta_j$ ).<sup>11</sup>

We provide existence and uniqueness results for different classes of equilibria in different parameter regions. Table 3.1 summarizes the results. Its first column corresponds to the numbered regions in Figure 3.2, the next two columns provide qualitative descriptions of the respective regions, the fourth column describes PNE and the last one illustrates them with an example. To facilitate comparisons between the regions, the same simple example is analyzed: six players, four of whom (referred to as "the majority") prefer action 1 and two ("the minority") prefer action 0. Equilibria in brackets exist under additional conditions.

<sup>11</sup>Note that the equilibrium networks described in brackets are more specific and relate equilibrium network structures to corresponding equilibrium action profiles.

The depicted sets of equilibria are not intended to be exhaustive for this particular example, but rather illustrative of possible classes of equilibria in each region.

Turning to equilibrium analysis, the first thing to note is that if the linking cost is very high (region 5), then  $S^{PNE}$  consists of all strategy profiles in  $\bar{S}$  that induce an empty network.

**Proposition 1** [EMPTY NETWORK]

Let  $c > \max\{\delta + \lambda, 1 - \delta + \lambda\}$ . A strategy profile  $s \in S^{PNE}$  if and only if  $p_{ij} = 0 \forall i, j \in N$ .

Such a high linking cost makes any link unprofitable. At the same time, an isolated player (not linked to anyone) gets zero utility regardless of the action she chooses, that is why any action profile is possible in equilibrium.

For all other parameter regions, let us first formulate necessary conditions for pairwise Nash equilibria. These conditions will differ for games of strategic complements ( $\delta \geq \frac{1}{2}$ ) and for games of strategic substitutes ( $\delta \leq \frac{1}{2}$ ).

**Proposition 2** [NECESSARY CONDITIONS FOR A PNE]

Let  $c < \max\{\delta + \lambda, 1 - \delta + \lambda\}$  and  $s \in S^{PNE}$ .

(i) If  $\delta \geq \frac{1}{2}$ , then for any  $i, j \in N$   $x_i = x_j$  implies  $p_{ij}p_{ji} = 1$ .

(ii) If  $\delta \leq \frac{1}{2}$ , then for any  $i, j \in N$   $x_i \neq x_j$  implies  $p_{ij}p_{ji} = 1$ .

Hence, if interactional incentives are such that players get higher utility from the links with matching actions than from the links with mismatching actions, then all players playing the same action must be linked in equilibrium. If interactional incentives are the contrary, then all pairs of players playing different actions must be linked. This result, although intuitive, is not trivial, as the linking cost might still outweigh the benefits for many pairs of players (see Figure 3.2). On the other hand, it is a very important result, as together with Proposition 1 and the symmetry of the setting (ex ante, players differ only with respect to their action preferences) it already pins down six classes of equilibrium network structures described above as an exhaustive list.

Contrary to the region 5 with its whole variety of equilibrium action profiles, the regions 4, 9 and 10 demonstrate another extreme: a unique equilibrium action profile here is the fully satisfying one – such that coincides with the preference profile. Moreover, in each of these regions an induced equilibrium network is also unique, which results into a unique PNE.

**Proposition 3** [UNIQUE PNE: FULLY SATISFYING ACTION PROFILE]

(i) Let  $\max\{\delta, 1 - \delta + \lambda\} < c < \delta + \lambda$ . A strategy profile  $s \in S^{PNE}$  if and only if  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \Leftrightarrow \theta_i = \theta_j$ .

	<i>Linking cost</i>	<i>Game</i>	<i>PNE</i>	<i>Example: <math>\theta = (1, 1, 1, 1, 0, 0)</math></i>
1	very low	coordination games	homogeneous action profile, complete network <i>if sufficient minority:</i> fully satisfying action profile, complete network	
2	low		homogeneous action profile, complete network <i>if sufficient minority:</i> fully satisfying action profile, complete network <i>if many players:</i> heterogeneous not fully satisfying action profile, two disconnected action cliques; <i>if many players and small minority:</i> heterogeneous not fully satisfying action profile, two partially connected action cliques	
3	medium		homogeneous action profile, complete network fully satisfying action profile, two disconnected action cliques <i>if many players:</i> heterogeneous not fully satisfying action profile, two disconnected action cliques	
4	high	games of strategic complements	fully satisfying action profile, two disconnected action cliques	
5	very high	all	any action profile, empty network	
6	very low	anti-coordination games	<i>if sufficient minority:</i> fully satisfying action profile, complete network <i>if small minority or few players:</i> heterogeneous not fully satisfying action profile, complete network	
7	low		<i>if sufficient minority:</i> fully satisfying action profile, complete network; heterogeneous not fully satisfying action profile, complete bipartite network (partition by action) <i>if many players and sufficient but not too large minority:</i> heterogeneous not fully satisfying action profile, union of a complete bipartite network and a clique	
8	medium		fully satisfying action profile, complete bipartite network (partition by action) <i>if many players:</i> heterogeneous not fully satisfying action profile, complete bipartite network (partition by action)	
9	high	games of strategic substitutes	fully satisfying action profile, complete bipartite network (partition by action)	
10	very low low medium	dominant action games	fully satisfying action profile, complete network	

Table 3.1: Pairwise Nash equilibria. Coloured numbers next to network nodes denote players' action preferences, colours of the nodes denote their actions (green – action 1, yellow – action 0).

(ii) Let  $\max\{1 - \delta, \delta + \lambda\} < c < 1 - \delta + \lambda$ . A strategy profile  $s \in S^{PNE}$  if and only if  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \Leftrightarrow \theta_i \neq \theta_j$ .

(iii) Let  $\lambda > |2\delta - 1|$  and  $c < \min\{\delta + \lambda, 1 - \delta + \lambda\}$ . A strategy profile  $s \in S^{PNE}$  if and only if  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \forall i, j \in N$ .

The intuition is the following. In the regions 4 and 9 (parts (i) and (ii) of the proposition respectively) the linking cost is not too high to exclude any possibility of a profitable link, but sufficiently high to prevent all except the most desirable types of links. Therefore, all players follow their action preferences and form links according to the interactional incentives (action matching in the region 4, or action mismatching in the region 9). In the region 10 (part (iii) of the proposition) the linking cost is lower, which permits links between players playing the same action as well as between those playing different actions. However, this region is characterized by strong action preferences ( $\lambda > |2\delta - 1|$ ), which leads to a unique, fully satisfying equilibrium action profile.

In fact, these are the only regions in which a PNE is always unique. In particular, in the regions 1, 2 and 3 (coordination games with at most medium linking cost) there always exist at least two PNE – complete networks with a homogeneous action profile.

**Proposition 4** [HOMOGENEOUS ACTION PROFILE]

Let  $\delta > \frac{1+\lambda}{2}$  and  $c < \delta$ . If  $x_i = x_j \forall i, j \in N$  and  $p_{ij} = 1 \forall i, j \in N$ , then  $s \in S^{PNE}$ .

In these regions links between players who play the same action are profitable regardless of whether these are their preferred actions or not. At the same time, coordination incentives secure homogeneous action profiles against unilateral action deviations.

The next three propositions concern other types of equilibria that can exist in these regions and provide sufficient conditions for their existence.

**Proposition 5** [FULLY SATISFYING ACTION PROFILE IN COORDINATION GAMES]

Let  $\delta > \frac{1+\lambda}{2}$ . If either of the conditions

$$(i) \quad c < 1 - \delta \text{ and } n^\pi \leq \frac{2\delta - 1 + \lambda}{2(2\delta - 1)}(n - 1) \forall \pi \in \{0, 1\},$$

$$(ii) \quad 1 - \delta < c < 1 - \delta + \lambda \text{ and } n^\pi \leq \frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 1) \forall \pi \in \{0, 1\}, \text{ or}$$

$$(iii) \quad 1 - \delta + \lambda < c < \delta$$

holds, then there exists  $s \in S^{PNE}$  with  $x_i = \theta_i \forall i \in N$ .

It can be shown (see the proof in the appendix) that for low and very low values of the linking cost a fully satisfying PNE inducing a complete network might exist (under additional conditions on the sizes of preferences groups), and for a medium linking cost there exists a



fully satisfying PNE inducing two disconnected cliques corresponding to different actions. Table 3.1 illustrates these possibilities (see the third equilibrium for each respective region).

In the regions 1 and 2, a sufficient (and, in fact, also necessary) condition for existence of a fully satisfying equilibrium is relative balancedness of the players' preference profile, that is, the preference majority must not be too large.<sup>12</sup> Note that as  $\delta$  approaches  $\frac{1+\lambda}{2}$ , the conditions on relative sizes of preference groups become less stringent (the respective ratios tend to 1), due to the growing weight of idiosyncratic utility component relative to its interactional component.

The following proposition concerns another equilibrium network structure – a network consisting of two disjoint disconnected cliques.

**Proposition 6** [TWO DISCONNECTED ACTION CLIQUES]

Let  $\delta > \frac{1+\lambda}{2}$ . If either of the conditions

- (i)  $1 - \delta < c < 1 - \delta + \lambda$  and  $n \geq 2 \left\lceil \frac{3\delta-1-c}{2\delta-1-\lambda} + 1 \right\rceil$ , or
- (ii)  $1 - \delta + \lambda < c < \delta$

holds, then there exists  $s \in S^{PNE}$  s.t.  $\forall i, j \in N$   $p_{ij} = 1 \Leftrightarrow x_i = x_j$  and  $\exists i, j \in N$  with  $p_{ij} = 0$ .

Due to coordination incentives of the game, these two disjoint cliques correspond to two different actions (see Proposition 2). Note that the two cliques are necessarily distinct, that is, not just a complete network with a homogeneous action profile as in Proposition 4. In the region 2 (part (i) of the proposition), an additional condition for existence of such a PNE is a sufficiently large number of players, as then the sizes of both action cliques can be sufficiently large to guarantee that pairwise deviations would be unprofitable.

Note that although this proposition concerns the regions 2 and 3, the same class of equilibria exists in the region 4 (Proposition 3, part (i)), where it is a unique equilibrium.

Finally, Proposition 7 provides sufficient conditions for existence of the last possible class of equilibria for coordination games – a network consisting of two disjoint partially connected cliques.

**Proposition 7** [TWO PARTIALLY CONNECTED ACTION CLIQUES]

Let  $\delta > \frac{1+\lambda}{2}$  and  $1 - \delta < c < 1 - \delta + \lambda$ . If  $n^\pi < \min\left\{\frac{\delta+\lambda-c}{3\delta-1-c}(n-1) - 3, n - 4 - \frac{\delta+\lambda-c}{2\delta-1-\lambda}\right\}$  for some  $\pi \in \{0, 1\}$ , then there exists  $s \in S^{PNE}$  s.t.  $\forall i, j \in N$   $x_i = x_j \Rightarrow p_{ij} = 1$ ,  $\exists i, j \in N$  s.t.  $x_i \neq x_j$  and  $p_{ij} = 1$ , and  $\exists k, l \in N$  s.t.  $x_k \neq x_l$  and  $p_{kl} = 0$ .

---

<sup>12</sup>Note that  $\frac{2\delta-1+\lambda}{2(2\delta-1)} \in [\frac{1}{2}, 1)$  (attaining the boundary value of  $\frac{1}{2}$  when  $\lambda = 0$ ) and  $\frac{\delta+\lambda-c}{3\delta-1-c} \in (\frac{1}{2}, 1)$ . In particular, when  $\lambda = 0$ , there is no equilibrium with a fully satisfying action profile in the region 1 (and the region 2 is empty).

Again, the two cliques correspond to two different actions, but now they are connected by at least one link. However, they are not fully connected, that is, the network is not complete. Such an equilibrium network exists in the region 2 if the number of players is sufficiently large (so that  $\min\{\frac{\delta+\lambda-c}{3\delta-1-c}(n-1)-3, n-4-\frac{\delta+\lambda-c}{2\delta-1-\lambda}\} > 2$ , as  $n^\pi \geq 2 \forall \pi \in \{0, 1\}$ ) and the preference minority is sufficiently small. This condition is sufficient but not necessary, however: in the example in Table 3.1 for parameter values  $\lambda = 0.1$ ,  $\delta = 0.7$  and  $c = 0.35$  this condition is not satisfied, even though such an equilibrium exists (the last depicted equilibrium for the region 2, where thick lines indicate the links between two action cliques).

Let us now turn to the remaining regions 6, 7 and 8 – anti-coordination games with at most medium linking cost. The following three propositions describe classes of equilibria possible there and provide sufficient conditions for their existence.

**Proposition 8** [FULLY SATISFYING ACTION PROFILE IN ANTI-COORDINATION GAMES]

Let  $\delta < \frac{1-\lambda}{2}$ . If either of the conditions

(i)  $c < \delta$  and  $n^\pi \geq \frac{1-2\delta-\lambda}{2(1-2\delta)}(n-1) \forall \pi \in \{0, 1\}$ ,

(ii)  $\delta < c < \delta + \lambda$  and  $n^\pi \geq \frac{1-2\delta-\lambda}{2-3\delta-c}(n-1) \forall \pi \in \{0, 1\}$ , or

(iii)  $\delta + \lambda < c < 1 - \delta$ ,

holds, then there exists  $s \in S^{PNE}$  with  $x_i = \theta_i \forall i \in N$ .

As the regions 6, 7 and 8 are symmetric to the regions 1, 2 and 3, respectively, fully satisfying equilibria there exist under similar conditions. For low or very low linking cost – the regions 6 and 7 – this condition is relative balancedness of the players’ preference profile (the preference minority must not be too small).<sup>13</sup> If a fully satisfying PNE exists there, it induces a complete network (and hence, is unique). In the region 8, corresponding to the medium linking cost, there always exists a unique fully satisfying PNE, which induces a complete bipartite network with the bipartition  $\{N^0, N^1\}$  (see Table 3.1).

The next proposition concerns more general complete bipartite networks as possible equilibrium network structures in anti-coordination games.

**Proposition 9** [COMPLETE BIPARTITE NETWORK]

Let  $\delta < \frac{1-\lambda}{2}$ . If either of the conditions

(i)  $\delta < c < \delta + \lambda$  and  $n^\pi > \frac{1-\delta+\lambda-c}{1-2\delta-\lambda} \forall \pi \in \{0, 1\}$ , or

---

<sup>13</sup>Note that  $\frac{1-2\delta-\lambda}{2(1-2\delta)} \in (0, \frac{1}{2}]$  (attaining the boundary value of  $\frac{1}{2}$  when  $\lambda = 0$ ) and  $\frac{1-2\delta-\lambda}{2-3\delta-c} \in (0, \frac{1}{2})$ . Here, even when  $\lambda = 0$ , the existence of a fully satisfying equilibrium for a very low linking cost is not completely excluded (unlike coordination games – see footnote 12), since anti-coordination incentives favour action heterogeneity.

(ii)  $\delta + \lambda < c < 1 - \delta$

holds, then there exists  $s \in S^{PNE}$  s.t.  $\forall i, j \in N$   $p_{ij} = 1 \Leftrightarrow x_i \neq x_j$ .

Note that in these bipartite networks players are partitioned according to their actions. As we already know from Proposition 2, the links between players choosing different actions constitute the minimal set of links for all games with strategic substitutes. This proposition shows that for some of these games – namely, for the regions 7 and 8 (for the region 9 see part (ii) of Proposition 3) – this set of links can also be maximal.

Finally, the last proposition demonstrates the last possible class of equilibrium network structures, that under some additional conditions on the sizes of preference groups is possible for games from the region 7.

**Proposition 10** [UNION OF A COMPLETE BIPARTITE NETWORK AND A CLIQUE]

Let  $\delta < \frac{1-\lambda}{2}$  and  $\delta < c < \delta + \lambda$ . If  $\frac{1-\delta+\lambda-c}{1-2\delta-\lambda} < \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} n^\pi \leq \min\{\frac{1-\delta+\lambda-c}{2-3\delta-c} n - 2, n - n^\pi - 2\}$  for some  $\pi \in \{0, 1\}$ , then there exists  $s \in S^{PNE}$  s.t.  $\forall i, j \in N$   $x_i \neq x_j \Rightarrow p_{ij} = 1$ ,  $\exists i, j \in N$  s.t.  $x_i = x_j$  and  $p_{ij} = 1$ , and  $\exists k, l \in N$  s.t.  $x_k = x_l$  and  $p_{kl} = 0$ .

Compared to a complete bipartite network from Proposition 9, this equilibrium network has additional links between some players from the same bipartition class (i.e. those playing the same action), but it is still less connected than a complete network. In particular, only players choosing their preferred actions can afford to have additional links (in Table 3.1, thick lines in the last two equilibrium networks for the region 7 indicate these additional links). The symmetry of the setting implies that all such players will be linked to each other, which derives the union of the complete bipartite network with the clique of all players choosing their preferred actions.

A sufficient number of players and a sufficient but not too large preference minority guarantees existence of this type of an equilibrium. This condition is not a necessary one, however: in the example in Table 3.1 for parameter values  $\lambda = 0.1$ ,  $\delta = 0.3$  and  $c = 0.35$ , the last two equilibria for the region 7 exist, even though the condition of Proposition 10 is not satisfied.

### 3.4 Discussion and conclusions

#### The role of players' heterogeneity in action preferences

There are classes of pairwise Nash equilibria that exist only for  $\lambda > 0$ . These are, in particular, the two classes with heterogeneous but not fully satisfying action profiles and incomplete asymmetric network structures (either two partially connected cliques or a union of a complete bipartite network and a clique). If  $\lambda = 0$  then the regions 2 and 7, which might give rise

to such equilibria, are empty. Hence, the most irregular equilibrium structures that can be achieved are due to players' heterogeneous action preferences.

Another class of equilibria that, generically, exists only for  $\lambda > 0$  is a fully satisfying action profile on a complete network. It might exist for up to medium values of the linking cost (in the regions 1, 2, 6, 7 or 10) and requires either a sufficiently balanced preference profile or relatively strong action preferences. If  $\lambda = 0$ , the equilibrium network there will still be complete, but the action profile will be either homogeneous (for coordination games) or heterogeneous but, generically, not fully satisfying (for anti-coordination games).<sup>14</sup>

### Alternative equilibrium concepts

Obviously, pairwise deviations defined in this paper do not cover the whole range of deviations that a pair of players can implement. One natural refinement of the pairwise Nash equilibrium concept would be to consider equilibria proof against *all* possible bilateral deviations.<sup>15</sup>

**Definition 2.** A strategy profile  $s$  is a *bilateral equilibrium (BE)* of the above game if it is a Nash equilibrium and for any pair of players  $i, j \in N$  and any strategy pair  $s'_i \in S_i, s'_j \in S_j$ ,

$$u_i(s'_i, s'_j, s_{-i-j}) > u_i(s) \Rightarrow u_j(s'_i, s'_j, s_{-i-j}) < u_j(s).$$

Since  $S^{BE} \subseteq S^{PNE}$ , no new equilibria can be derived if we consider this alternative equilibrium concept. What would be interesting is to verify if all the classes of PNE that we described can survive these more stringent equilibrium requirements. Although a complete characterization of bilateral equilibria lies outside the scope of this paper, let us make one important observation: irregular equilibrium structures are still possible under the BE equilibrium concept (the following figure provides an illustration).

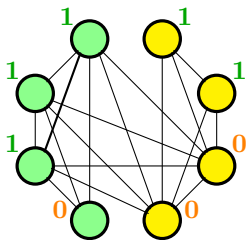


Figure 3.3: A bilateral equilibrium of the coordination game with  $\delta = 0.7$ ,  $\lambda = 0.1$  and  $c = 0.35$ . Coloured numbers denote players' action preferences, colours of the nodes – players' actions in this equilibrium (green corresponds to action 1, yellow – to action 0).

<sup>14</sup>Equilibria in the case of anti-coordination games with homogeneous players are characterized in Bramoullé (2007): when the network is complete, the proportion of agents playing a strategy approximately equals the mixed equilibrium probability of this strategy.

<sup>15</sup>The following definition is adopted from Goyal and Vega-Redondo (2007), with the only difference that in this paper an action is also a part of a strategy.

One could further refine the bilateral equilibrium set by considering deviations by coalitions consisting of more than two players. Dutta and Mutuswami (1997) and Jackson and van den Nouweland (2005) introduced the notion of strong stability of a network that refers to a situation where no coalition of players can rearrange their links to achieve a strong (or even weak – in the latter paper) improvement. These notions can be adapted to our framework with a simultaneous action choice. It would be interesting to verify if the conclusion of Jackson and van den Nouweland (2005) that strongly stable networks coincide with the set of efficient networks holds also in our framework.

### Efficiency

A natural next step in our research is the efficiency analysis of possible outcomes. Figure 3.4 depicts the regions of parameter values corresponding to different sets of efficient strategy profiles. These regions partition the regions of different equilibrium sets in Figure 3.2.

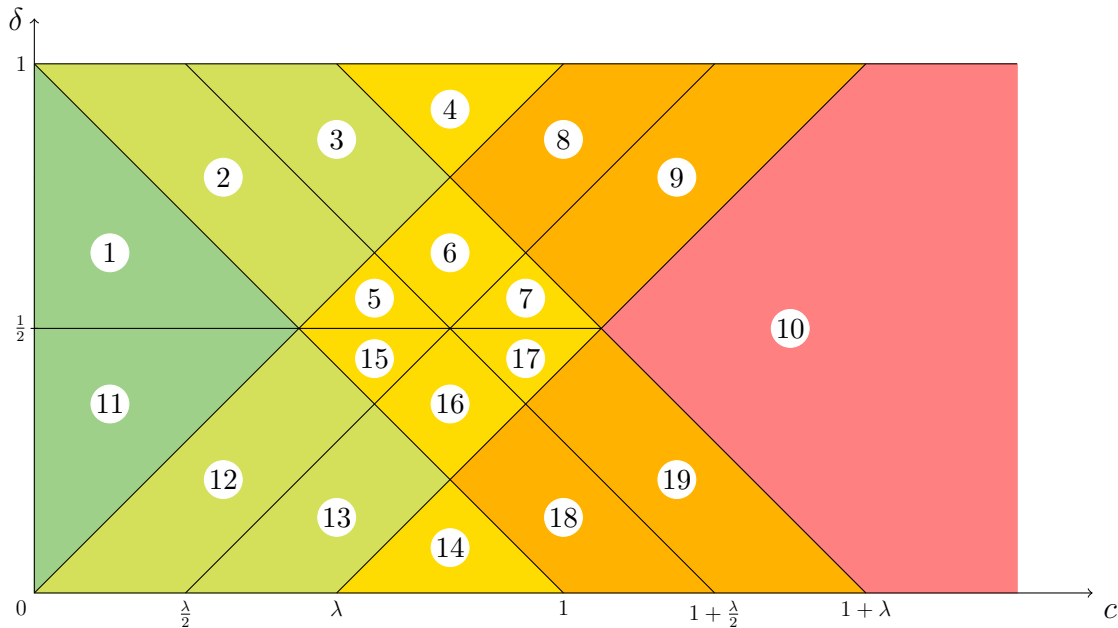


Figure 3.4: Regions of the linking cost values that correspond to different efficient networks. Each horizontal section defines a game  $(\delta, \lambda)$ .

One way to derive efficient strategy profiles is to maximize aggregate welfare over all created links (which is equivalent to maximizing aggregate welfare over all players). Altogether, six types of links are possible: three types that connect players from the same preference group and three types that connect players with different action preferences (see Table 3.2). Aggregate welfare is then the sum of corresponding link payoffs over all links.

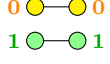
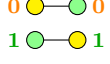
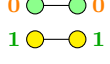

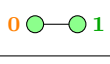
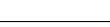
	Link type	Link payoff	Number of such links
1		$2(\delta + \lambda - c)$	$\frac{m^0(m^0-1)}{2} + \frac{m^1(m^1-1)}{2}$
2		$2(1 - \delta - c) + \lambda$	$m^0(n^0 - m^0) + m^1(n^1 - m^1)$
3		$2(\delta - c)$	$\frac{(n^0-m^0)(n^0-m^0-1)}{2} + \frac{(n^1-m^1)(n^1-m^1-1)}{2}$
4		$2(1 - \delta + \lambda - c)$	$m^0m^1$
5		$2(\delta - c) + \lambda$	$m^0(n^1 - m^1) + m^1(n^0 - m^0)$
6		$2(1 - \delta - c)$	$(n^0 - m^0)(n^1 - m^1)$

Table 3.2: Six possible types of links, their contributions to aggregate welfare and respective quantities. Here  $m^\pi = |\{i \in N^\pi \mid x_i = \theta_i\}|$  for  $\pi \in \{0, 1\}$ .

Each region in Figure 3.4 defines which types of links contribute to aggregate welfare (generate positive payoffs) and hence can be present in efficient profiles in this region. For instance, in the region 1 all links are profitable, which implies that efficient networks here are necessarily complete. Summing up link payoffs over all links and subtracting the constant part of aggregate welfare derives the following welfare maximization problem for this region:

$$\max_{\substack{m^0 \in \{0, \dots, n^0\} \\ m^1 \in \{0, \dots, n^1\}}} 2(2\delta - 1)(m^0 - m^1)^2 + 2(2\delta - 1)(n^1 - n^0)(m^0 - m^1) + \lambda(n - 1)(m^0 + m^1),$$

which is equivalent to

$$\max_{\substack{x, y \in \mathbb{Z}_+ \\ x+y \leq 2n^0 \\ x-y \leq 2n^1}} 2(2\delta - 1)y^2 + 2(2\delta - 1)(n^1 - n^0)y + \lambda(n - 1)x.$$

For other regions, the maximization problems are derived in a similar way. After efficient strategy profiles are characterized, we can compare them with equilibrium profiles and find when achieving efficiency is guaranteed, when it is possible and when not. We leave these questions for future research.

## Appendix to Chapter 3

### Proof of Lemma 1

Take  $s \in S^{NE}$  and suppose that  $s \notin \bar{S}$ . The latter implies that  $\exists i, j \in N$  s.t.  $p_{ij} = 1$  and  $p_{ji} = 0$ . Consider  $s'_i = (x'_i, p'_i)$  with  $x'_i = x_i$ ,  $p'_{ij} = 0$  and  $p'_{ik} = p_{ik} \forall k \neq j$ . Then  $u_i(s'_i, s_{-i}) = u_i(s) + \varepsilon > u_i(s)$ , i.e.  $i$ 's payoff is strictly higher with such a strategy deviation, and hence  $s \notin S^{NE}$ . By contradiction we proved that  $s \in S^{NE}$  implies  $s \in \bar{S}$ .  $\square$

### Proof of Lemma 2

Necessity follows trivially, so let us prove sufficiency. Let  $s = (x, p) \in S$  be such that  $u_i(s'_i, s_{-i}) \leq u_i(s) \forall i \in N \forall s'_i \in \bar{S}_i(s_{-i})$ . If  $p_{ij} = 1 \forall i, j \in N$ , then  $\bar{S}_i(s_{-i}) = S_i \forall i \in N$ , and hence the proof is completed. Suppose  $\exists i, j \in N$  s.t.  $p_{ji} = 0$ , that is,  $S_i \setminus \bar{S}_i(s_{-i}) \neq \emptyset$ . Take  $s''_i = (x''_i, p''_i) \in S_i \setminus \bar{S}_i(s_{-i})$  and let us prove that  $u_i(s''_i, s_{-i}) \leq u_i(s)$ . Denote  $J = \{j \in N \setminus \{i\} : p''_{ij} = 1 \text{ and } p_{ji} = 0\}$  and consider now  $s'_i = (x'_i, p'_i)$  such that  $x'_i = x''_i$ ,  $p'_{ij} = 0 \forall j \in J$  and  $p'_{ij} = p''_{ij} \forall j \notin J \cup \{i\}$ . Then  $u_i(s''_i, s_{-i}) < u_i(s'_i, s_{-i})$ , which together with  $s'_i \in \bar{S}_i(s_{-i})$  implies  $u_i(s''_i, s_{-i}) < u_i(s)$ . Hence,  $s$  is a Nash equilibrium.  $\square$

### Proof of Lemma 3

Let us first note that Assumption 1 implies that  $u_{ij}(x_i, x_j) \neq 0 \forall i, j \in N$ .

Necessity. Let  $s \in S^{PNE}$  and pick  $i, j \in N$  s.t.  $p_{ij} = 1$ . According to Lemma 1,  $p_{ji} = p_{ij} = 1$ . Without loss of generality, suppose  $u_{ij}(x_i, x_j) < 0$ . Consider now  $s'_i = (x'_i, p'_i) \in S_i$  s.t.  $x'_i = x_i$ ,  $p'_{ij} = 0$  and  $p'_{ik} = p_{ik} \forall k \neq j$ . Then  $u_i(s'_i, s_{-i}) > u_i(s)$ , and hence  $s \notin S^{PNE}$ .

Sufficiency. Let  $s \in S^{PNE}$  and  $i, j \in N$  be s.t. both  $u_{ij}(x_i, x_j) > 0$  and  $u_{ji}(x_j, x_i) > 0$ . Suppose,  $p_{ij} = 0$ . According to Lemma 1,  $p_{ji} = p_{ij} = 0$ . Consider a pairwise deviation  $s'_i, s'_j$  s.t.  $x'_i = x_i$ ,  $x'_j = x_j$ ,  $p'_{ij} = p'_{ji} = 1$  and  $p'_{lk} = p_{lk} \forall l \in \{i, j\} \forall k \notin \{i, j\}$ . Then  $u_i(s'_i, s'_j, s_{-i-j}) > u_i(s)$  and  $u_j(s'_i, s'_j, s_{-i-j}) > u_j(s)$ , and hence  $s \notin S^{PNE}$ .  $\square$

### Proof of Lemma 4

Necessity follows trivially, so let us prove sufficiency. Consider  $s \in \bar{S}^+$  and let  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s) \forall i \in N$ . We need to prove that  $u_i(s'_i, s_{-i}) \leq u_i(s) \forall s'_i \in \bar{S}_i(s_{-i}) \forall i \in N$  (see Lemma 2). Fix any  $i \in N$ . If  $s'_i = (x'_i, p'_i)$  is such that  $x'_i = x_i$  and  $p'_i \neq p_i$ , then  $s \in \bar{S}^+$  implies  $u_i(s'_i, s_{-i}) < u_i(s)$ . If  $s'_i$  is such that  $x'_i \neq x_i$ , then  $x'_i = \tilde{x}_i$ , and hence  $u_i(s'_i, s_{-i}) \leq u_i(\tilde{s}_i, s_{-i}) = \sum_{j: u_{ij}(\tilde{x}_i, x_j) > 0} p_{ij} p_{ji} u_{ij}(\tilde{x}_i, x_j) \leq u_i(s)$ . Therefore,  $u_i(s'_i, s_{-i}) \leq u_i(s) \forall s'_i \in \bar{S}_i(s_{-i}) \forall i \in N$ .  $\square$

### Proof of Lemma 5

Necessity follows from Lemma 3 and the definition of a PNE. Let us prove sufficiency. Take  $s \in \bar{S}^+$  and let conditions of the lemma hold for all  $i \in N$  and for all  $j \in N$  s.t.  $p_{ij} p_{ji} = 0$ . According to Lemma 4, condition (1) implies that  $s \in S^{NE}$ . We are left to prove that for all  $i, j \in N$  s.t.  $p_{ij} p_{ji} = 0$ ,  $s$  is proof against the pairwise deviation  $((x_i, p'_i), (x_j, p'_j))$ .

Note that since  $s \in \bar{S}^+$ ,  $p_{ij} p_{ji} = 0$  together with Assumption 1 implies that either  $u_{ij}(x_i, x_j) < 0$  or  $u_{ji}(x_j, x_i) < 0$ . Consequently, either  $u_i((x_i, p'_i), (x_j, p'_j), s_{-i-j}) < u_i(s)$  or  $u_j((x_i, p'_i), (x_j, p'_j), s_{-i-j}) < u_j(s)$  respectively. Hence,  $s$  is also proof against all possible pairwise deviations, i.e.  $s \in S^{PNE}$ .  $\square$

## Proof of Lemma 6

Necessity. Let  $s \in S^{PNE}$ . Lemma 3 implies that  $s \in \bar{S}^+$ . It follows from the definition of a PNE that condition (1) holds for all  $i \in N$ . To prove the necessity of condition (2), consider arbitrary  $i, j \in N$  with  $p_{ij}p_{ji} = 0$ . If  $u_{ji}(\hat{x}_i, x_j) > 0$ , then  $u_j((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) = u_j(s) + u_{ji}(\hat{x}_i, x_j) > u_j(s)$ , and condition (2) follows then from the definition of a PNE. Let now  $u_{ji}(\hat{x}_i, x_j) < 0$  and assume that condition (2) does not hold, i.e.  $u_i(s) \leq u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) = u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) + u_{ij}(\hat{x}_i, x_j) \leq u_i(s) + u_{ij}(\hat{x}_i, x_j)$ , which together with Assumption 1 implies  $u_{ij}(\hat{x}_i, x_j) > 0$ . The rest of the proof shows that in this case  $s \notin S^{PNE}$ .

Note that  $u_{ji}(\hat{x}_i, x_j) < 0$  together with  $u_{ij}(\hat{x}_i, x_j) > 0$  is only possible if  $x_j \neq \theta_j$  and  $\hat{x}_i = \theta_i$ . Hence,  $x_i \neq \theta_i$ , and consequently,  $u_{ij}(x_i, x_j) = u_{ji}(x_i, x_j) = \begin{cases} \delta & \text{if } \theta_i = \theta_j \\ 1 - \delta & \text{if } \theta_i \neq \theta_j \end{cases}$ .

Since  $s \in \bar{S}^+$  and  $p_{ij}p_{ji} = 0$ , it must be that  $u_{ij}(x_i, x_j) = u_{ji}(x_i, x_j) < 0$ . Noting additionally that  $u_{ji}(\hat{x}_i, x_j) = 1 - u_{ji}(x_i, x_j) < 0$  derives  $\max\{\delta, 1 - \delta\} < c$ . It must also be that  $c < \max\{\delta + \lambda, 1 - \delta + \lambda\}$ , since otherwise  $u_{ij}(\hat{x}_i, x_j) > 0$  is violated. Finally, taking into account that  $c \notin C_{\delta, \lambda}^h$ , we derive  $\max\{\delta, 1 - \delta\} < c < \min\{\delta + \lambda, 1 - \delta + \lambda\}$ . But then  $u_{kl}(x_k, x_l) > 0$  if and only if  $x_k = \theta_k$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{kl} = 1 \Leftrightarrow x_k = \theta_k \wedge x_l = \theta_l\}$ . Above we derived that both  $x_i \neq \theta_i$  and  $x_j \neq \theta_j$ , which implies  $p_i = p_j = 0$ , and thus  $u_i(s) = u_j(s) = 0$ . However, a pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = \theta_i$  and  $x'_j = \theta_j$  would be Pareto improving for  $i$  and  $j$ :  $u_i(s'_i, s'_j, s_{-i-j}) = u_{ij}(x'_i, x'_j) > 0 = u_i(s)$  and  $u_j(s'_i, s'_j, s_{-i-j}) = u_{ji}(x'_i, x'_j) > 0 = u_j(s)$ . This contradicts that  $s \in S^{PNE}$ . Thus, condition (2) is also necessary for  $s \in S^{PNE}$ .

Sufficiency. Take  $s \in \bar{S}^+$  and let conditions of the lemma hold for all  $i \in N$  and for all  $j \in N$  s.t.  $p_{ij}p_{ji} = 0$ . According to Lemma 4, condition (1) implies that  $s \in S^{NE}$ . We are left to prove that for all  $i, j \in N$  s.t.  $p_{ij}p_{ji} = 0$ ,  $s$  is proof against two pairwise deviations:  $((x_i, \hat{p}_i), (x_j, \hat{p}_j))$  and  $((\hat{x}_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j))$  with  $\hat{x}_i \neq x_i$  and  $\hat{x}_j \neq x_j$ . For the first one, see the analogous proof of Lemma 5. Consider now the deviation  $((\hat{x}_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j))$ .

*Case 1:*  $c > \max\{\delta + \lambda, 1 - \delta + \lambda\}$ . Then  $u_{kl}(x_k, x_l) < 0 \forall k, l \in N$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{kl} = 0 \forall k, l \in N\}$ , which implies  $u_i(s) = u_j(s) = 0$ . Since then  $u_i((\hat{x}_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j), s_{-i-j}) < 0 = u_i(s)$ ,  $s$  is proof against the deviation  $((\hat{x}_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j))$ .

*Case 2:*  $c < \max\{\delta + \lambda, 1 - \delta + \lambda\} = \delta + \lambda$ . Then  $\delta \geq \frac{1}{2}$ . Taking into account that  $c \notin C_{\delta, \lambda}^h$ , it must be that  $c < \max\{\delta, 1 - \delta + \lambda\}$ . First, let  $x_i \neq x_j$ . Then  $1 - \delta \leq \delta$  implies  $u_{ij}(\hat{x}_i, \hat{x}_j) \leq u_{ij}(\hat{x}_i, x_j)$ , and hence, due to condition (2) of the lemma,  $u_i((\hat{x}_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j), s_{-i-j}) = u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) - u_{ij}(\hat{x}_i, x_j) + u_{ij}(\hat{x}_i, \hat{x}_j) < u_i(s)$ , i.e.  $s$  is proof against this deviation. Second, let  $x_i = x_j$ . We show that this leads to a contradiction. Note that it must be that  $c > \delta$ , as otherwise both  $u_{ij}(x_i, x_j) > 0$  and  $u_{ji}(x_i, x_j) > 0$ , which together with  $p_{ij}p_{ji} = 0$  contradicts  $s \in \bar{S}^+$ . Then  $c < \max\{\delta, 1 - \delta + \lambda\}$  implies  $c < 1 - \delta + \lambda$ . Note also that either  $x_i \neq \theta_i$  or  $x_j \neq \theta_j$  (or both), as otherwise  $u_{ij}(x_i, x_j) = u_{ji}(x_i, x_j) =$



$\delta + \lambda - c > 0$ . Without loss of generality, let  $x_i \neq \theta_i$ . Since  $s \in \bar{S}^+$  and  $c > \delta \geq 1 - \delta$ , it follows that  $p_i = 0$ . But then  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) = u_{ij}(\hat{x}_i, x_j) = 1 - \delta + \lambda - c > 0 = u_i(s)$ , which contradicts condition (2) of the lemma.

*Case 3:*  $c < \max\{\delta + \lambda, 1 - \delta + \lambda\} = 1 - \delta + \lambda$ . Then  $\delta \leq \frac{1}{2}$ . Taking into account that  $c \notin C_{\delta, \lambda}^h$ , it must be that  $c < \max\{\delta + \lambda, 1 - \delta\}$ . The rest of the proof is symmetric to the previous case. First, let  $x_i = x_j$ . Then  $\delta \leq 1 - \delta$  implies  $u_{ij}(\hat{x}_i, \hat{x}_j) \leq u_{ij}(\hat{x}_i, x_j)$ , and hence, due to condition (2) of the lemma,  $u_i((\hat{x}_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j), s_{-i-j}) = u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) - u_{ij}(\hat{x}_i, x_j) + u_{ij}(\hat{x}_i, \hat{x}_j) < u_i(s)$ , i.e.  $s$  is proof against this deviation. Second, we show that  $x_i \neq x_j$  is impossible, as it leads to a contradiction. Let  $x_i \neq x_j$ . Note that it must be that  $c > 1 - \delta$ , as otherwise both  $u_{ij}(x_i, x_j) > 0$  and  $u_{ji}(x_i, x_j) > 0$ . Hence,  $c < \delta + \lambda$ . Note also that, like in the previous case, either  $x_i \neq \theta_i$  or  $x_j \neq \theta_j$  (or both). Without loss of generality, let  $x_i \neq \theta_i$ , and consequently,  $p_i = 0$ . But then  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) = u_{ij}(\hat{x}_i, x_j) = \delta + \lambda - c > 0 = u_i(s)$ , which contradicts condition (2) of the lemma.

In all cases  $s$  is proof against the deviation  $((\hat{x}_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j))$ , and thus  $s \in S^{PNE}$ .  $\square$

### Proof of Proposition 1

Let  $c > \max\{\delta + \lambda, 1 - \delta + \lambda\}$ . Since  $\lambda \geq 0$ , it means that  $u_{ij}(x_i, x_j) < 0 \forall x_i, x_j \in \{0, 1\}$ . Hence,  $\bar{S}^+ = \{s \in S \mid p_{ij} = 0 \forall i, j \in N\}$ . According to Lemma 3,  $S^{PNE} \subseteq \bar{S}^+$ , which proves necessity of  $p_{ij} = 0 \forall i, j \in N$  for every  $s \in S^{PNE}$ . To prove its sufficiency, we can use Lemma 5. First, observe that  $\forall i \in N \forall s \in \bar{S}^+ u_i(s) = 0$ , and hence  $u_i(\tilde{s}_i, s_{-i}) = u_i(s)$ . Second,  $\forall i, j \in N \forall s \in \bar{S}^+ u_i((x'_i, p'_i), (x'_j, p'_j), s_{-i-j}) < 0 = u_i(s)$  for all admissible  $x'_i$  and  $x'_j$ . Hence, every  $s \in \bar{S}^+$  is proof against all admissible deviations, which implies  $s \in S^{PNE}$ .  $\square$

### Proof of Proposition 2

Let  $s \in S^{PNE}$  and  $c < \max\{\delta + \lambda, 1 - \delta + \lambda\}$ . Lemma 1 implies  $p_{ij} = p_{ji} \forall i, j \in N$ .

(i) Let  $\delta \geq \frac{1}{2}$ , which implies  $c < \delta + \lambda$ . Suppose  $\exists i, j \in N$  s.t.  $x_i = x_j$  but  $p_{ij} = p_{ji} = 0$ .

We show that in each of the following cases (that cover all possibilities) there exists a profitable pairwise deviation, which contradicts  $s \in S^{PNE}$ .

*Case 1:*  $c < \delta$ . Consider a pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = x_i$  and  $x'_j = x_j$ .  $u_i(s'_i, s'_j, s_{-i-j}) = u_i(s) + u_{ij}(x_i, x_j) > u_i(s)$ , as  $u_{ij}(x_i, x_j) \geq \delta - c > 0$ , and similarly,  $u_j(s'_i, s'_j, s_{-i-j}) > u_j(s)$ .

*Case 2:*  $x_i = \theta_i$  and  $x_j = \theta_j$ . Again, a pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = x_i$  and  $x'_j = x_j$  is profitable for  $i$  and  $j$ :  $u_i(s'_i, s'_j, s_{-i-j}) > u_i(s)$  and  $u_j(s'_i, s'_j, s_{-i-j}) > u_j(s)$ , as  $u_{ij}(x_i, x_j) = u_{ji}(x_i, x_j) = \delta + \lambda - c > 0$ .

*Case 3:*  $\delta < c < 1 - \delta + \lambda$  and either  $x_i \neq \theta_i$  or  $x_j \neq \theta_j$  (or both). Without loss of generality, let  $x_i \neq \theta_i$ . Since  $c > \max\{\delta, 1 - \delta\}$ , Lemma 3 implies  $p_i = 0$ . Consider a

pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = \theta_i$  and  $x'_j = \theta_j$ :  $u_i(s'_i, s'_j, s_{-i-j}) = u_{ij}(x'_i, x'_j) > 0 = u_i(s)$ , as  $c < \min\{\delta + \lambda, 1 - \delta + \lambda\}$ , and similarly,  $u_j(s'_i, s'_j, s_{-i-j}) = u_j(s) + u_{ji}(x'_i, x'_j) > u_j(s)$  (note that the last equality holds both for  $x_j = \theta_j$  and for  $x_j \neq \theta_j$ , as in the latter case  $p_j = 0$ ).

*Case 4:*  $c > \max\{\delta, 1 - \delta + \lambda\}$  and either  $x_i \neq \theta_i$  or  $x_j \neq \theta_j$  (or both). Without loss of generality, let  $x_i \neq \theta_i$ . As in the previous case, Lemma 3 implies  $p_i = 0$ . There are two possibilities: either  $\theta_i = \theta_j$  or  $\theta_i \neq \theta_j$ . Consider the first possibility. Then a pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = \theta_i$  and  $x'_j = \theta_j$  is profitable for  $i$  and  $j$ :  $u_i(s'_i, s'_j, s_{-i-j}) = u_{ij}(x'_i, x'_j) = \delta + \lambda - c > 0 = u_i(s)$  and similarly,  $u_j(s'_i, s'_j, s_{-i-j}) = u_j(s) + u_{ji}(x'_i, x'_j) > u_j(s)$  (note that the last equality holds both for  $x_j = \theta_j$  and for  $x_j \neq \theta_j$ , as in the latter case  $p_j = 0$ ). Consider the second possibility,  $\theta_i \neq \theta_j$ . Take a third player  $k$  with  $\theta_k = \theta_i$ . As  $p_i = 0$ , a pairwise deviation  $(s'_i, s'_k)$  with  $x'_i = \theta_i$  and  $x'_k = \theta_k$  is feasible and, moreover, profitable for  $i$  and  $k$  (see the above reasoning for  $i$  and  $j$ ).

As in each case there exists a profitable pairwise deviation, our assumption contradicts  $s \in S^{PNE}$ . Hence, it must be that  $\forall i, j \in N$   $x_i = x_j$  implies  $p_{ij} = p_{ji} = 1$ .

- (ii) Let  $\delta \leq \frac{1}{2}$ , which implies  $c < 1 - \delta + \lambda$ . Suppose  $\exists i, j \in N$  s.t.  $x_i \neq x_j$  but  $p_{ij} = p_{ji} = 0$ . We show that in each of the following cases (that cover all possibilities) there exists a profitable pairwise deviation, which contradicts  $s \in S^{PNE}$ . The cases are symmetric to those in part (i), and hence the rest of the proof is analogous to the proof of (i).

*Case 1:*  $c < 1 - \delta$ . A pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = x_i$  and  $x'_j = x_j$  is profitable for  $i$  and  $j$ , as  $u_{ij}(x_i, x_j) \geq 1 - \delta - c > 0$  and  $u_{ji}(x_i, x_j) \geq 1 - \delta - c > 0$ .

*Case 2:*  $x_i = \theta_i$  and  $x_j = \theta_j$ . Again, a pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = x_i$  and  $x'_j = x_j$  is profitable for  $i$  and  $j$ , as  $u_{ij}(x_i, x_j) = u_{ji}(x_i, x_j) = 1 - \delta + \lambda - c > 0$ .

*Case 3:*  $1 - \delta < c < \delta + \lambda$  and either  $x_i \neq \theta_i$  or  $x_j \neq \theta_j$  (or both). Without loss of generality, let  $x_i \neq \theta_i$ , which implies  $p_i = 0$ . A pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = \theta_i$  and  $x'_j = \theta_j$  is profitable for  $i$  and  $j$ :  $u_i(s'_i, s'_j, s_{-i-j}) = u_{ij}(x'_i, x'_j) > 0 = u_i(s)$  and  $u_j(s'_i, s'_j, s_{-i-j}) = u_j(s) + u_{ji}(x'_i, x'_j) > u_j(s)$  (the last equality holds both for  $x_j = \theta_j$  and for  $x_j \neq \theta_j$ ).

*Case 4:*  $c > \max\{\delta + \lambda, 1 - \delta\}$  and either  $x_i \neq \theta_i$  or  $x_j \neq \theta_j$  (or both). Without loss of generality, let  $x_i \neq \theta_i$ , which implies  $p_i = 0$ . If  $\theta_i \neq \theta_j$ , then a pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = \theta_i$  and  $x'_j = \theta_j$  is profitable for  $i$  and  $j$ , as  $u_{ij}(x'_i, x'_j) = u_{ji}(x'_i, x'_j) = 1 - \delta + \lambda - c > 0$ . If  $\theta_i = \theta_j$ , then there must be a player  $k$  with  $\theta_k \neq \theta_i$ , and a pairwise deviation  $(s'_i, s'_k)$  with  $x'_i = \theta_i$  and  $x'_k = \theta_k$  is profitable for  $i$  and  $k$ , as  $u_{ik}(x'_i, x'_k) = u_{ki}(x'_i, x'_k) = 1 - \delta + \lambda - c > 0$ .

In each case our assumption contradicts  $s \in S^{PNE}$ . Thus, it must be that  $\forall i, j \in N$   $x_i \neq x_j$  implies  $p_{ij} = p_{ji} = 1$ .

□

### Proof of Proposition 3

- (i) Let  $\max\{\delta, 1 - \delta + \lambda\} < c < \delta + \lambda$ . Since  $\lambda \geq 0$ , it means that  $u_{ij}(x_i, x_j) > 0$  if and only if  $\theta_i = x_i = x_j$ , otherwise  $u_{ij}(x_i, x_j) < 0$ . Hence,  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow \theta_i = x_i = x_j = \theta_j\}$ .

Necessity. Consider a strategy profile  $s \in S^{PNE}$ , and suppose that  $x_i \neq \theta_i$  for some  $i \in N$ . Note that  $S^{PNE} \subseteq \bar{S}^+$  (see Lemma 3). Then it must be that  $p_{ij} = 0 \forall j \in N$ . Take another player  $j$  with  $\theta_j = \theta_i$  (such a player must exist, as  $n^\pi \geq 2 \forall \pi \in \{0, 1\}$ ). A pairwise deviation  $(s'_i, s'_j)$  with  $x'_i = \theta_i$  and  $x'_j = \theta_j$  is Pareto improving for  $i$  and  $j$ :  $u_i(s'_i, s'_j, s_{-i-j}) = u_{ij}(x'_i, x'_j) > 0 = u_i(s)$  and  $u_j(s'_i, s'_j, s_{-i-j}) = u_j(s) + u_{ji}(x'_i, x'_j) > u_j(s)$  (note that the last equality holds both for  $x_j = \theta_j$  and for  $x_j \neq \theta_j$ , as in the latter case  $p_j = 0$ ). Hence,  $s \notin S^{PNE}$ .

Now let  $s \in S^{PNE}$  and  $x_i = \theta_i \forall i \in N$ . Since  $S^{PNE} \subseteq \bar{S}^+$ , it must be that  $p_{ij} = 1 \Leftrightarrow \theta_i = \theta_j$ , which completes this part of the proof.

Sufficiency. Consider a strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \Leftrightarrow \theta_i = \theta_j$ . Then  $s \in \bar{S}^+$ . It suffices to verify that conditions of Lemma 5 hold. First, fix a player  $i$ . Since  $u_{ij}(\tilde{x}_i, x_j) < 0 \forall x_j \in \{0, 1\}$ , it must be that  $\tilde{p}_{ij} = 0 \forall j \in N$ , and hence  $u_i(\tilde{s}_i, s_{-i}) = 0 \leq u_i(s)$ . Second, fix a pair of players  $i$  and  $j$  s.t.  $p_{ij}p_{ji} = 0$ . Consider a pairwise deviation  $(s'_i, s'_j)$  with either  $x'_i \neq x_i$  or  $x'_j \neq x_j$ . Without loss of generality, let  $x'_i \neq x_i$ . Since  $x'_i \neq \theta_i$  implies  $u_{ik}(x'_i, x_k) < 0 \forall x_k \in \{0, 1\}$ , it follows that  $u_i(s'_i, s'_j, s_{-i-j}) = u_i((x'_i, p_i), s_{-i}) + u_{ij}(x'_i, x'_j) \leq u_i(s) + u_{ij}(x'_i, x'_j) < u_i(s)$ . Hence,  $s$  is proof against all admissible deviations, which implies  $s \in S^{PNE}$ .

- (ii) Let  $\max\{1 - \delta, \delta + \lambda\} < c < 1 - \delta + \lambda$ . Then  $u_{ij}(x_i, x_j) > 0$  if and only if  $\theta_i = x_i \neq x_j$ , and hence,  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow \theta_i = x_i \neq x_j = \theta_j\}$ . The rest of the proof is identical to that of part (i) with the only difference: the proof of necessity of  $x_i = \theta_i \forall i \in N$  for  $s \in S^{PNE}$  uses a pairwise deviation of  $i$  and  $j$  s.t.  $\theta_j \neq \theta_i$ .

- (iii) Let  $\lambda > |2\delta - 1|$  and  $c < \min\{\delta + \lambda, 1 - \delta + \lambda\}$ . The first inequality is equivalent to  $\frac{1-\lambda}{2} < \delta < \frac{1+\lambda}{2}$ , which implies  $\max\{\delta, 1 - \delta\} < \min\{\delta + \lambda, 1 - \delta + \lambda\}$ . Several cases have to be considered separately:

*Case 1:*  $c < \min\{\delta, 1 - \delta\}$ . Then  $u_{ij}(x_i, x_j) > 0 \forall x_i, x_j \in \{0, 1\}$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \forall i, j \in N\}$ .

*Case 2:*  $1 - \delta < c < \delta$ . Then  $u_{ij}(x_i, x_j) > 0$  if and only if either  $x_i = \theta_i$  or  $x_i = x_j$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i = x_j)\}$ .

*Case 3:*  $\delta < c < 1 - \delta$ . Then  $u_{ij}(x_i, x_j) > 0$  if and only if either  $x_i = \theta_i$  or  $x_i \neq x_j$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i \neq x_j)\}$ .

*Case 4:*  $c > \max\{\delta, 1 - \delta\}$ . Then  $u_{ij}(x_i, x_j) > 0$  if and only if  $x_i = \theta_i$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow (x_i = \theta_i \wedge x_j = \theta_j)\}$ .

Necessity. Consider a strategy profile  $s \in S^{PNE}$ , and suppose that  $x_i \neq \theta_i$  for some  $i \in N$ . Lemma 3 implies that  $s \in \bar{S}^+$ . First, consider cases 1, 2 and 3. Let us prove that in each of these cases  $\exists j \in N$  s.t.  $p_{ij}p_{ji} = 1$ . Suppose not, that is  $p_{ij}p_{ji} = 0 \forall j \in N$ . Getting a contradiction in case 1 is trivial. In cases 2 and 3 any pairwise deviation  $(s'_i, s'_j)$  s.t.  $x'_i = \theta_i$  and  $x'_j = \theta_j$  is Pareto improving for  $i$  and  $j$ :  $u_i(s'_i, s'_j, s_{-i-j}) = u_{ij}(x'_i, x'_j) > 0 = u_i(s)$  and  $u_j(s'_i, s'_j, s_{-i-j}) \geq u_j(s) + u_{ji}(x'_i, x'_j) > u_j(s)$ . Note that the penultimate inequality holds as equality for  $x_j = \theta_j$ , and if  $x_j \neq \theta_j$  then for all  $k \in N$  s.t.  $p_{jk}p_{kj} = 1$  (if they exist) it holds that either  $u_{jk}(x_j, x_k) = \delta - c < 1 - \delta + \lambda - c = u_{jk}(x'_j, x_k)$  (case 2) or  $u_{jk}(x_j, x_k) = 1 - \delta - c < \delta + \lambda - c = u_{jk}(x'_j, x_k)$  (case 3), and hence  $u_j(s) = \sum_{\{k \in N: p_{jk}p_{kj}=1\}} u_{jk}(x_j, x_k) \leq u_j((x'_j, p_j), s_{-j})$  (with equality if  $\nexists k \in N$  s.t.  $p_{jk}p_{kj} = 1$ ). Thus, we have proved that  $\exists j \in N$  s.t.  $p_{ij}p_{ji} = 1$ .

Now consider the unilateral deviation  $\tilde{s}_i$ . Since  $\tilde{x}_i = \theta_i$ ,  $u_{ij}(\tilde{x}_i, x_j) > 0 \forall j \in N$ , and hence  $\tilde{p}_{ij} = p_{ij}$ . Note also that for all  $j \in N$  s.t.  $p_{ij}p_{ji} = 1$  (above we have shown that at least one such  $j$  exists)  $u_{ij}(\tilde{x}_i, x_j) \geq \min\{\delta + \lambda - c, 1 - \delta + \lambda - c\} > \max\{\delta - c, 1 - \delta - c\} \geq u_{ij}(x_i, x_j)$ . Then  $u_i(\tilde{s}_i, s_{-i}) = u_i((\tilde{x}_i, p_i), s_{-i}) = \sum_{\{j \in N: p_{ij}p_{ji}=1\}} u_{ij}(\tilde{x}_i, x_j) > \sum_{\{j \in N: p_{ij}p_{ji}=1\}} u_{ij}(x_i, x_j) = u_i(s)$ , i.e.  $\tilde{s}_i$  is a payoff-improving deviation. Hence,  $s \notin S^{PNE}$ . By contradiction, we have proved that  $s \in S^{PNE}$  implies  $x_i = \theta_i \forall i \in N$  in cases 1, 2 and 3.

Consider case 4. Since  $s \in \bar{S}^+$  and  $x_i \neq \theta_i$ , it must be that  $p_{ij} = 0 \forall j \in N$ . But then a pairwise deviation  $(s'_i, s'_j)$  s.t.  $x'_i = \theta_i$  and  $x'_j = \theta_j$  is Pareto improving for  $i$  and  $j$ :  $u_i(s'_i, s'_j, s_{-i-j}) = u_{ij}(x'_i, x'_j) > 0 = u_i(s)$  and  $u_j(s'_i, s'_j, s_{-i-j}) = u_j(s) + u_{ji}(x'_i, x'_j) > u_j(s)$ . Note that the last equality holds both for  $x_j = \theta_j$  and for  $x_j \neq \theta_j$ , as in the latter case  $u_j(s) = 0$ . Hence, we have proved that  $s \in S^{PNE}$  must have  $x_i = \theta_i \forall i \in N$  also in case 4.

Now let  $s \in S^{PNE}$  and  $x_i = \theta_i \forall i \in N$ . Since  $s \in \bar{S}^+$ , it must be that  $p_{ij} = 1 \forall i, j \in N$  (in each of the four cases), which completes this part of the proof.

Sufficiency. Consider a strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \forall i, j \in N$ . Then  $s \in \bar{S}^+$  (in each of the four cases). As there are no feasible pairwise deviations, it suffices to verify that the first condition of Lemma 5 holds. Fix a player  $i$ . Note that for all  $j \in N$  it holds that  $u_{ij}(x_i, x_j) \geq \min\{\delta + \lambda - c, 1 - \delta + \lambda - c\} > \max\{\delta - c, 1 - \delta - c\} \geq$

$u_{ij}(\tilde{x}_i, x_j)$ . Then  $u_i(\tilde{s}_i, s_{-i}) = \sum_{\{j \in N: \tilde{p}_{ij} p_{ji} = 1\}} u_{ij}(\tilde{x}_i, x_j) = \sum_{j \in N} \max\{u_{ij}(\tilde{x}_i, x_j), 0\} < \sum_{j \in N} u_{ij}(x_i, x_j) = u_i(s)$ , and hence, according to Lemma 5,  $s \in S^{PNE}$ .

□

### Proof of Proposition 4

Let  $\delta > \frac{1+\lambda}{2}$  and  $c < \delta$  and consider a strategy profile with  $x_i = x_j \forall i, j \in N$  and  $p_{ij} = 1 \forall i, j \in N$ . For all  $i, j \in N$  either  $u_{ij}(x_i, x_j) = \delta - c > 0$  or  $u_{ij}(x_i, x_j) = \delta + \lambda - c > 0$ , hence  $s \in \bar{S}^+$ . We can now apply Lemma 5.

First, take a player with  $x_i = \theta_i$ . Then  $u_i(s) = (\delta + \lambda - c)(n - 1)$ , while with the unilateral deviation  $\tilde{s}_i$  she gets  $u_i((\tilde{x}_i, \tilde{p}_i), s_{-i}) = \sum_{\{j \in N: j \neq i\}} \max\{1 - \delta - c, 0\} \leq (1 - \delta - c)(n - 1) < (\delta + \lambda - c)(n - 1) = u_i(s)$  (the last inequality follows from  $\delta > \frac{1+\lambda}{2}$ , which is equivalent to  $1 - \delta + \lambda < \delta$ ).

Now, take a player with  $x_i \neq \theta_i$ . Then  $u_i(s) = (\delta - c)(n - 1)$ , while with the unilateral deviation  $\tilde{s}_i$  she gets  $u_i((\tilde{x}_i, \tilde{p}_i), s_{-i}) = \sum_{\{j \in N: j \neq i\}} \max\{1 - \delta + \lambda - c, 0\} \leq (1 - \delta + \lambda - c)(n - 1) < (\delta - c)(n - 1) = u_i(s)$ .

In either case, such a unilateral deviation is unprofitable for  $i$ . And since there are no feasible pairwise deviations, Lemma 5 implies that  $s \in S^{PNE}$ . □

### Proof of Proposition 5

Note that if any of the conditions (i), (ii) or (iii) holds, then  $c \notin C_{\delta, \lambda}^h$ , and hence Lemma 6 is applicable here.

- (i) Let  $\delta > \frac{1+\lambda}{2}$ ,  $c < 1 - \delta$  and  $n^\pi \leq \frac{2\delta - 1 + \lambda}{2(2\delta - 1)}(n - 1) \forall \pi \in \{0, 1\}$ . We will prove that the strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \forall i, j \in N$  is a PNE.

Note that  $u_{ij}(x_i, x_j) > 0 \forall x_i, x_j \in \{0, 1\}$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \forall i, j \in N\}$ . Obviously,  $s \in \bar{S}^+$ . There are no possible pairwise deviations from  $s$ , hence it suffices to show that  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s) \forall i \in N$  (see Lemma 6). Without loss of generality, take a player with  $\theta_i = 0$ . Then  $u_i(s) = (\delta + \lambda - c)(n^0 - 1) + (1 - \delta + \lambda - c)n^1$ , and  $u_i((\tilde{x}_i, \tilde{p}_i), s_{-i}) = u_i((\tilde{x}_i, p_i), s_{-i}) = (1 - \delta - c)(n^0 - 1) + (\delta - c)n^1$ . Consequently,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $(1 - \delta)(n^0 - 1) + \delta n^1 \leq (\delta + \lambda)(n^0 - 1) + (1 - \delta + \lambda)n^1$ . Rearranging terms and substituting  $n^0$  for  $n - n^1$ , we can get an equivalent inequality:  $n^1 \leq \frac{2\delta - 1 + \lambda}{2(2\delta - 1)}(n - 1)$ . Similarly, for a player with  $\theta_i = 1$ ,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $n^0 \leq \frac{2\delta - 1 + \lambda}{2(2\delta - 1)}(n - 1)$ . In either case, such a unilateral deviation is unprofitable for  $i$ , and hence  $s \in S^{PNE}$ .

- (ii) Let  $\delta > \frac{1+\lambda}{2}$ ,  $1 - \delta < c < 1 - \delta + \lambda$  and  $n^\pi \leq \frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 1) \forall \pi \in \{0, 1\}$ . We will prove that the strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \forall i, j \in N$  is a PNE.

Note that  $u_{ij}(x_i, x_j) > 0$  if and only if either  $x_i = \theta_i$  or  $x_i = x_j$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i = x_j)\}$ . Then  $s \in \bar{S}^+$ . The rest of the proof is similar to the proof of part (i): it suffices to show that  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s) \forall i \in N$ . Take a player with  $\theta_i = 0$ . Then  $u_i(s) = (\delta + \lambda - c)(n^0 - 1) + (1 - \delta + \lambda - c)n^1$ , and  $u_i((\tilde{x}_i, \tilde{p}_i), s_{-i}) = \sum_{\{j \in N: x_j = \tilde{x}_i\}} u_{ij}(\tilde{x}_i, x_j) = (\delta - c)n^1$ . Hence,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $\delta n^1 \leq (\delta + \lambda - c)(n^0 - 1) + (1 - \delta + \lambda)n^1$ , or equivalently,  $n^1 \leq \frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 1)$ . Similarly, for a player with  $\theta_i = 1$ ,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $n^0 \leq \frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 1)$ . In either case, such a unilateral deviation is unprofitable for  $i$ , and hence  $s \in S^{PNE}$ .

- (iii) Let  $\delta > \frac{1+\lambda}{2}$  and  $1 - \delta + \lambda < c < \delta$ . We will prove that the strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \Leftrightarrow \theta_i = \theta_j$  is a PNE.

Here  $u_{ij}(x_i, x_j) > 0$  if and only if  $x_i = x_j$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow x_i = x_j\}$ . Note that  $s \in \bar{S}^+$ . As in the previous two parts of the proof, we can apply Lemma 6. First, take a player with action preference  $\theta_i$  and consider the unilateral deviation  $\tilde{s}_i$ . Then  $u_i(s) = (\delta + \lambda - c)(n^{\theta_i} - 1) \geq 0 = u_i(\tilde{s}_i, s_{-i})$ , which implies that  $\tilde{s}_i$  is unprofitable. Second, take two unlinked players  $i$  and  $j$  (with respective action preferences  $\theta_i \neq \theta_j$ ) and consider their pairwise deviation  $((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j))$ . Since  $n^\pi \geq 2 \forall \pi \in \{0, 1\}$ , we can derive:  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) = (1 - \delta - c)(n^{\theta_i} - 1) + (\delta - c) < \delta - c < (\delta + \lambda - c)(n^{\theta_i} - 1) = u_i(s)$ . Hence, according to Lemma 6,  $s \in S^{PNE}$ .

□

## Proof of Proposition 6

- (i) Let  $1 - \delta < c < 1 - \delta + \lambda < \delta$  (the last inequality is equivalent to  $\delta > \frac{1+\lambda}{2}$ ). Then  $u_{ij}(x_i, x_j) > 0$  if and only if either  $x_i = \theta_i$  or  $x_i = x_j$ . Hence,  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i = x_j)\}$ .

Let  $n \geq 2 \lceil \frac{3\delta - 1 - c}{2\delta - 1 - \lambda} + 1 \rceil$ . Then it must be that  $n^\pi \geq \lceil \frac{3\delta - 1 - c}{2\delta - 1 - \lambda} + 1 \rceil$  for some  $\pi \in \{0, 1\}$ . Without loss of generality, let  $n^0 \geq \lceil \frac{3\delta - 1 - c}{2\delta - 1 - \lambda} + 1 \rceil$  and consider such a strategy profile  $s$  that  $|\{i \in N \mid x_i = 1\}| = |\{i \in N^0 \mid x_i = 1\}| = \lceil \frac{3\delta - 1 - c}{2\delta - 1 - \lambda} + 1 \rceil$  and  $\forall i, j \in N$   $p_{ij} = 1 \Leftrightarrow x_i = x_j$ . Then  $a^\pi := |\{i \in N \mid x_i = \pi\}| \geq \lceil \frac{3\delta - 1 - c}{2\delta - 1 - \lambda} + 1 \rceil \forall \pi \in \{0, 1\}$ . Note that  $s \in \bar{S}^+$  (as  $x_i = \theta_i$  and  $x_j = \theta_j$  implies  $x_i = x_j = 0$ ), and let us apply Lemma 6.

First, fix a player  $i$ . If  $\tilde{x}_i = \theta_i$ , then  $u_{ij}(\tilde{x}_i, x_j) = 1 - \delta + \lambda - c > 0 \forall j$  s.t.  $p_{ij} = 1$ , hence  $\tilde{p}_i = p_i$  and  $u_i(\tilde{s}_i, s_{-i}) = \sum_{\{j \in N: p_{ij}=1\}} (1 - \delta + \lambda - c) < \sum_{\{j \in N: p_{ij}=1\}} (\delta - c) = u_i(s)$ . If  $\tilde{x}_i \neq \theta_i$ , then  $u_{ij}(\tilde{x}_i, x_j) = 1 - \delta - c < 0 \forall j$  s.t.  $p_{ij} = 1$ , hence  $\tilde{p}_i = 0$  and  $u_i(\tilde{s}_i, s_{-i}) = 0 \leq u_i(s)$ .

Second, fix a pair of players  $i$  and  $j$  s.t.  $p_{ij} = 0$  (i.e.  $x_i \neq x_j$ ) and consider a pairwise deviation  $((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j))$ . If  $\hat{x}_i = \theta_i$ , then  $u_i(s) = \sum_{\{j \in N: p_{ij}=1\}} (\delta - c) =$

$(a^{1-\theta_i} - 1)(\delta - c)$ , while  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) = (a^{1-\theta_i} - 1)(1 - \delta + \lambda - c) + (\delta + \lambda - c)$ . Thus,  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) < u_i(s)$  if and only if  $(a^{1-\theta_i} - 1)(1 - \delta + \lambda - c) + (\delta + \lambda - c) < (a^{1-\theta_i} - 1)(\delta - c)$ , which is equivalent to  $a^{1-\theta_i} > \frac{3\delta - 1 - c}{2\delta - 1 - \lambda}$ . This last inequality holds true due to  $a^\pi \geq \lceil \frac{3\delta - 1 - c}{2\delta - 1 - \lambda} + 1 \rceil \forall \pi \in \{0, 1\}$ , and hence condition (2) of Lemma 6 is satisfied. Similarly, if  $\hat{x}_i \neq \theta_i$ , then  $u_i(s) = (a^{1-\theta_i} - 1)(\delta + \lambda - c)$ , while  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) = (a^{1-\theta_i} - 1)(1 - \delta - c) + (\delta - c)$ . In this case,  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) < u_i(s)$  if and only if  $(a^{1-\theta_i} - 1)(1 - \delta - c) + (\delta - c) < (a^{1-\theta_i} - 1)(\delta + \lambda - c)$ , which is equivalent to  $a^{1-\theta_i} > \frac{3\delta - 1 + \lambda - c}{2\delta - 1 + \lambda}$ . However, this last inequality is weaker than the respective one in the previous case (using  $c < \delta$ , one can show that  $\frac{3\delta - 1 + \lambda - c}{2\delta - 1 + \lambda} < \frac{3\delta - 1 - c}{2\delta - 1 - \lambda}$ ). Hence, condition (2) of Lemma 6 is satisfied also in this case, and we can conclude that  $s \in S^{PNE}$ .

- (ii) Let  $1 - \delta + \lambda < c < \delta$ . Then the strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \Leftrightarrow x_i = x_j$  is a PNE (see the proof of part (iii) of Proposition 5).

□

### Proof of Proposition 7

Let  $1 - \delta < c < 1 - \delta + \lambda < \delta$  (the last inequality is equivalent to  $\delta > \frac{1 + \lambda}{2}$ ). Then  $u_{ij}(x_i, x_j) > 0$  if and only if either  $x_i = \theta_i$  or  $x_i = x_j$ , and  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i = x_j)\}$ . Without loss of generality, let  $n^0 < \min\{\frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 1) - 3, n - 4 - \frac{\delta + \lambda - c}{2\delta - 1 - \lambda}\}$ .

Let us introduce some additional notation,  $a^\pi := |\{i \in N \mid x_i = \pi\}|$  for  $\pi \in \{0, 1\}$ , and consider such a strategy profile  $s$  that  $x_i = 0 \forall i \in N^0$ ,  $|\{i \in N^1 \mid x_i = 1\}| = \lceil \frac{2\delta - 1 - \lambda}{\delta + \lambda - c} n^0 \rceil + 2$  and  $p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i = x_j)$ . Note that here  $a^1 = |\{i \in N^1 \mid x_i = 1\}|$ . First of all, we show that  $0 < a^1 < n^1$ , and hence that  $s$  indeed induces two partially connected action cliques:  $\exists i, j \in N$  s.t.  $\theta_i = 0 = x_i \neq x_j = 1 = \theta_j$  and  $p_{ij} = 1$ , and  $\exists k, l \in N$  s.t.  $\theta_k \neq 0 = x_k \neq x_l = 1 = \theta_l$  and  $p_{kl} = 0$ .

Thus, we need to prove that  $0 < \lceil \frac{2\delta - 1 - \lambda}{\delta + \lambda - c} n^0 \rceil + 2 < n^1$ . As  $\frac{2\delta - 1 - \lambda}{\delta + \lambda - c} \in (0, 1)$ , the first inequality is obvious, while the second one can be derived from  $n^0 < \frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 1) - 3$  in several steps: using that  $\frac{\delta + \lambda - c}{3\delta - 1 - c} \in (\frac{1}{2}, 1)$ , we derive  $n^0 < \frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 3)$ , or equivalently,  $\frac{3\delta - 1 - c}{\delta + \lambda - c} n^0 + 1 < n - 2$ , which implies  $\lceil \frac{3\delta - 1 - c}{\delta + \lambda - c} n^0 \rceil < n - 2$ , or equivalently,  $\lceil (1 + \frac{2\delta - 1 - \lambda}{\delta + \lambda - c}) n^0 \rceil < n - 2$ , and hence  $\lceil \frac{2\delta - 1 - \lambda}{\delta + \lambda - c} n^0 \rceil < n - n^0 - 2 = n^1 - 2$ . Hence, the strategy profile  $s$  indeed induces two connected action cliques. The rest of the proof shows that  $s \in S^{PNE}$ .

We can apply Lemma 6. Note that  $s \in \bar{S}^+$  and fix a player  $i$ . First, let  $\theta_i = 0$ . Then  $u_i(s) = (\delta + \lambda - c)(a^0 - 1) + (1 - \delta + \lambda - c)a^1$  and  $u_i(\tilde{s}_i, s_{-i}) = (\delta - c)a^1$ . It follows that  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $a^1 \leq \frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 1)$ . Substituting  $a^1 = \lceil \frac{2\delta - 1 - \lambda}{\delta + \lambda - c} n^0 \rceil + 2$ , we get an equivalent expression:  $\lceil \frac{2\delta - 1 - \lambda}{\delta + \lambda - c} n^0 \rceil \leq \frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 1) - 2$ . This, however, is always true, as  $n^0 < \frac{\delta + \lambda - c}{3\delta - 1 - c}(n - 1) - 3$  and  $\frac{2\delta - 1 - \lambda}{\delta + \lambda - c} \in (0, 1)$ . Second, let  $\theta_i = x_i = 1$ . Then  $u_i(s) = (\delta + \lambda - c)(a^1 - 1) + (1 - \delta + \lambda - c)n^0$  and  $u_i(\tilde{s}_i, s_{-i}) = (\delta - c)n^0$ . Consequently,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $a^1 \geq \frac{2\delta - 1 - \lambda}{\delta + \lambda - c} n^0 + 1$ , i.e.  $\lceil \frac{2\delta - 1 - \lambda}{\delta + \lambda - c} n^0 \rceil + 2 \geq \frac{2\delta - 1 - \lambda}{\delta + \lambda - c} n^0 + 1$ ,

which obviously holds. Finally, let  $\theta_i = 1 \neq x_i$ . Then  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $(1 - \delta + \lambda - c)(a^0 - 1) \leq (\delta - c)(a^0 - 1)$ , which holds true due to  $\delta > \frac{1+\lambda}{2}$ .

Now, fix a pair of players  $i$  and  $j$  s.t.  $p_{ij} = 0$ . It must be that  $x_i \neq x_j$  and, without loss of generality,  $x_i \neq \theta_i$  and  $x_j = \theta_j$  (if also  $x_j \neq \theta_j$ , it would contradict  $x_i \neq x_j$ ). It is left to check that condition (2) of Lemma 6 holds for both  $i$  and  $j$ . For  $i$ ,  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) < u_i(s)$  if and only if  $(1 - \delta + \lambda - c)(a^0 - 1) + (\delta + \lambda - c) < (\delta - c)(a^0 - 1)$ , which is equivalent to  $a^1 < n - 1 - \frac{\delta + \lambda - c}{2\delta - 1 - \lambda}$ , or  $\lceil \frac{2\delta - 1 - \lambda}{\delta + \lambda - c} n^0 \rceil + 2 < n - 1 - \frac{\delta + \lambda - c}{2\delta - 1 - \lambda}$ . However, our assumption  $n^0 < n - 4 - \frac{\delta + \lambda - c}{2\delta - 1 - \lambda}$  together with  $\frac{2\delta - 1 - \lambda}{\delta + \lambda - c} \in (0, 1)$  makes it hold true. For  $j$ ,  $u_j((x_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j), s_{-i-j}) < u_j(s)$  if and only if  $(1 - \delta - c)(a^1 - 1) + (\delta - c)n^0 + (\delta - c) < (\delta + \lambda - c)(a^1 - 1) + (1 - \delta + \lambda - c)n^0$ , or equivalently,  $a^1 > \frac{2\delta - 1 - \lambda}{2\delta - 1 + \lambda}n^0 + \frac{\delta - c}{2\delta - 1 + \lambda} + 1$ , or  $\lceil \frac{2\delta - 1 - \lambda}{\delta + \lambda - c}n^0 \rceil + 2 > \frac{2\delta - 1 - \lambda}{2\delta - 1 + \lambda}n^0 + \frac{\delta - c}{2\delta - 1 + \lambda} + 1$ . Note, however, that  $\frac{\delta - c}{2\delta - 1 + \lambda} \in (0, 1)$  and  $\frac{2\delta - 1 - \lambda}{\delta + \lambda - c} > \frac{2\delta - 1 - \lambda}{2\delta - 1 + \lambda}$ , hence condition (2) holds also for  $j$  and, according to Lemma 6,  $s \in S^{PNE}$ .  $\square$

### Proof of Proposition 8

The proof is analogous to the proof of Proposition 5. Note that if any of the conditions (i), (ii) or (iii) holds, then  $c \notin C_{\delta, \lambda}^h$ , and hence Lemma 6 is applicable here.

- (i) Let  $\delta < \frac{1-\lambda}{2}$ ,  $c < \delta$  and  $n^\pi \geq \frac{1-2\delta-\lambda}{2(1-2\delta)}(n-1) \forall \pi \in \{0, 1\}$ . We will prove that the strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \forall i, j \in N$  is a PNE.

Note that  $u_{ij}(x_i, x_j) > 0 \forall x_i, x_j \in \{0, 1\}$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \forall i, j \in N\}$ . Obviously,  $s \in \bar{S}^+$ . There are no possible pairwise deviations from  $s$ , hence it suffices to show that  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s) \forall i \in N$  (Lemma 6). Without loss of generality, take a player with  $\theta_i = 0$ . Then, as in Proposition 5,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $(1 - \delta)(n^0 - 1) + \delta n^1 \leq (\delta + \lambda)(n^0 - 1) + (1 - \delta + \lambda)n^1$ . Rearranging terms and substituting  $n^0$  for  $n - n^1$ , we get an equivalent inequality:  $n^1 \geq \frac{1-2\delta-\lambda}{2(1-2\delta)}(n-1)$ . Similarly, for a player with  $\theta_i = 1$ ,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $n^0 \geq \frac{1-2\delta-\lambda}{2(1-2\delta)}(n-1)$ . In either case, such a unilateral deviation is unprofitable for  $i$ , and hence  $s \in S^{PNE}$ .

- (ii) Let  $\delta < \frac{1-\lambda}{2}$ ,  $\delta < c < \delta + \lambda$  and  $n^\pi \geq \frac{1-2\delta-\lambda}{2-3\delta-c}(n-1) \forall \pi \in \{0, 1\}$ . We will prove that the strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \forall i, j \in N$  is a PNE.

Note that  $u_{ij}(x_i, x_j) > 0$  if and only if either  $x_i = \theta_i$  or  $x_i \neq x_j$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i \neq x_j)\}$ . Then  $s \in \bar{S}^+$ . According to Lemma 6, it suffices to show that  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s) \forall i \in N$ . Take a player with  $\theta_i = 0$ . Then  $u_i(s) = (\delta + \lambda - c)(n^0 - 1) + (1 - \delta + \lambda - c)n^1$ , and  $u_i((\tilde{x}_i, \tilde{p}_i), s_{-i}) = \sum_{\{j \in N: x_j \neq \tilde{x}_i\}} u_{ij}(\tilde{x}_i, x_j) = (1 - \delta - c)(n^0 - 1)$ . Hence,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $(1 - \delta)(n^0 - 1) \leq (\delta + \lambda)(n^0 - 1) + (1 - \delta + \lambda - c)n^1$ , or equivalently,  $n^1 \geq \frac{1-2\delta-\lambda}{2-3\delta-c}(n-1)$ . Similarly, for a player with  $\theta_i = 1$ ,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $n^0 \geq \frac{1-2\delta-\lambda}{2-3\delta-c}(n-1)$ . In either case, such a unilateral deviation is unprofitable for  $i$ , and hence  $s \in S^{PNE}$ .



- (iii) Let  $\delta < \frac{1-\lambda}{2}$  and  $\delta + \lambda < c < 1 - \delta$ . We will prove that the strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \Leftrightarrow \theta_i \neq \theta_j$  is a PNE.

Here  $u_{ij}(x_i, x_j) > 0$  if and only if  $x_i \neq x_j$ , and hence  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow x_i \neq x_j\}$ . Note that  $s \in \bar{S}^+$  and let us check the other conditions of Lemma 6. First, take a player with action preference  $\theta_i$  and consider the unilateral deviation  $\tilde{s}_i$ . Then  $u_i(s) = (1 - \delta + \lambda - c)n^{1-\theta_i} \geq 0 = u_i(\tilde{s}_i, s_{-i})$ , which implies that  $\tilde{s}_i$  is unprofitable. Second, take two unlinked players  $i$  and  $j$  (with  $\theta_i = \theta_j$ ) and consider their pairwise deviation  $((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j))$ . Since  $n^\pi \geq 2 \forall \pi \in \{0, 1\}$ , we can derive:  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) = (\delta - c)n^{1-\theta_i} + (1 - \delta - c) < 1 - \delta - c < (1 - \delta + \lambda - c)n^{1-\theta_i} = u_i(s)$ . Hence, according to Lemma 6,  $s \in S^{PNE}$ .

□

### Proof of Proposition 9

- (i) Let  $\delta < c < \delta + \lambda < 1 - \delta$  (the last inequality is equivalent to  $\delta < \frac{1-\lambda}{2}$ ). Then  $u_{ij}(x_i, x_j) > 0$  if and only if either  $x_i = \theta_i$  or  $x_i \neq x_j$ . Hence,  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i \neq x_j)\}$ .

Let  $n^\pi > \frac{1-\delta+\lambda-c}{1-2\delta-\lambda} \forall \pi \in \{0, 1\}$ , and consider such a strategy profile  $s$  that  $x_i \neq \theta_i \forall i \in N$  and  $p_{ij} = 1 \Leftrightarrow x_i \neq x_j \forall i, j \in N$ . Note that  $s \in \bar{S}^+$  and let us apply Lemma 6. First, fix a player  $i$  with action preference  $\theta_i$ . Then  $u_i(\tilde{s}_i, s_{-i}) = (\delta + \lambda - c)n^{1-\theta_i} < (1 - \delta - c)n^{1-\theta_i} = u_i(s)$ . Second, fix a pair of players  $i$  and  $j$  s.t.  $p_{ij} = 0$  (i.e.  $x_i = x_j$ ). Then  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) < u_i(s)$  if and only if  $(\delta + \lambda - c)n^{1-\theta_i} + (1 - \delta + \lambda - c) < (1 - \delta - c)n^{1-\theta_i}$ , or equivalently,  $n^{1-\theta_i} > \frac{1-\delta+\lambda-c}{1-2\delta-\lambda}$ . Similarly,  $u_j((x_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j), s_{-i-j}) < u_j(s)$  if and only if  $n^{\theta_i} > \frac{1-\delta+\lambda-c}{1-2\delta-\lambda}$ . As all conditions of Lemma 6 are satisfied, we conclude that  $s \in S^{PNE}$ .

- (ii) Let  $\delta + \lambda < c < 1 - \delta$ . Then the strategy profile with  $x_i = \theta_i \forall i \in N$  and  $p_{ij} = 1 \Leftrightarrow x_i \neq x_j$  is a PNE (see the proof of part (iii) of Proposition 8).

□

### Proof of Proposition 10

Let  $\delta < c < \delta + \lambda < 1 - \delta$  (the last inequality is equivalent to  $\delta < \frac{1-\lambda}{2}$ ). Then  $u_{ij}(x_i, x_j) > 0$  if and only if either  $x_i = \theta_i$  or  $x_i \neq x_j$ , and  $\bar{S}^+ = \{s \in \bar{S} \mid p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i \neq x_j)\}$ . Without loss of generality, let  $\frac{1-\delta+\lambda-c}{1-2\delta-\lambda} < \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} n^0 \leq \min\{\frac{1-\delta+\lambda-c}{2-3\delta-c} n - 2, n - n^0 - 2\}$ .

Let us introduce some additional notation,  $a^\pi := |\{i \in N \mid x_i = \pi\}|$  for  $\pi \in \{0, 1\}$ , and consider such a strategy profile  $s$  that  $x_i = 0 \forall i \in N^0$ ,  $|\{i \in N^1 \mid x_i = 1\}| = \lceil \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) \rceil + 1$  and  $p_{ij} = 1 \Leftrightarrow ((x_i = \theta_i \wedge x_j = \theta_j) \vee x_i \neq x_j)$ . Note that here  $a^1 = |\{i \in N^1 \mid x_i = 1\}|$ .

As  $a^0 \geq n^0 \geq 2$ , there exist  $i, j \in N$  s.t.  $\theta_i = x_i = x_j = \theta_j = 0$  and  $p_{ij} = 1$ . Let us show that  $a^1 < n^1$ , and hence  $a^0 > n^0$ , implying that there exist also  $k, l \in N$  s.t.  $0 = \theta_k = x_k = x_l \neq \theta_l = 1$  and  $p_{kl} = 0$ . Noting that  $\frac{1-2\delta-\lambda}{1-\delta+\lambda-c} \in (0, 1)$ , we can derive  $\left\lceil \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) \right\rceil - 1 < \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) < \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} n^0 \leq n - n^0 - 2 = n^1 - 2$ , which is equivalent to  $\left\lceil \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) \right\rceil + 1 < n^1$  and is exactly what we wanted to show. The rest of the proof shows that  $s \in S^{PNE}$ .

We can apply Lemma 6. Note that  $s \in \bar{S}^+$  and fix a player  $i$ . First, let  $\theta_i = 0$ . Then  $u_i(s) = (1 - \delta + \lambda - c) a^1 + (\delta + \lambda - c)(n^0 - 1)$  and  $u_i(\tilde{s}_i, s_{-i}) = (1 - \delta - c)(n^0 - 1)$ . It follows that  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $a^1 \geq \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1)$ , which is always true, as  $a^1 = \left\lceil \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) \right\rceil + 1$ . Second, let  $\theta_i = x_i = 1$ . Then  $u_i(s) = (1 - \delta + \lambda - c) a^0 + (\delta + \lambda - c)(a^1 - 1)$  and  $u_i(\tilde{s}_i, s_{-i}) = (1 - \delta - c)(a^1 - 1)$ . Consequently,  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $a^1 \leq \frac{1-\delta+\lambda-c}{2-3\delta-c} n + \frac{1-2\delta-\lambda}{2-3\delta-c}$ , which also holds, as  $a^1 = \left\lceil \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) \right\rceil + 1 < \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) + 2 \leq \left( \frac{1-\delta+\lambda-c}{2-3\delta-c} n - 2 \right) - \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} + 2 < \frac{1-\delta+\lambda-c}{2-3\delta-c} n + \frac{1-2\delta-\lambda}{2-3\delta-c}$ . Finally, let  $\theta_i = 1 \neq x_i$ . Then  $u_i(\tilde{s}_i, s_{-i}) \leq u_i(s)$  if and only if  $(\delta + \lambda - c) a^1 \leq (1 - \delta - c) a^1$ , which holds true due to  $\delta < \frac{1-\lambda}{2}$ .

Now, fix a pair of players  $i$  and  $j$  s.t.  $p_{ij} = 0$ . It must be that  $x_i = x_j$  and, without loss of generality,  $x_i \neq \theta_i$  (if both  $x_i = \theta_i$  and  $x_j = \theta_j$ , it would contradict  $p_{ij} = 0$ ). It is left to check that condition (2) of Lemma 6 holds for both  $i$  and  $j$ . For  $i$ ,  $u_i((\hat{x}_i, \hat{p}_i), (x_j, \hat{p}_j), s_{-i-j}) < u_i(s)$  if and only if  $(\delta + \lambda - c) a^1 + (1 - \delta + \lambda - c) < (1 - \delta - c) a^1$ , which is equivalent to  $a^1 > \frac{1-\delta+\lambda-c}{1-2\delta-\lambda}$ . It holds true, as  $a^1 = \left\lceil \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) \right\rceil + 1 \geq \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) + 1 > \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} n^0 > \frac{1-\delta+\lambda-c}{1-2\delta-\lambda}$ . If  $x_j \neq \theta_j$ , then condition (2) for  $j$  coincides with the above one for  $i$ . If  $x_j = \theta_j$ , then  $u_j((x_i, \hat{p}_i), (\hat{x}_j, \hat{p}_j), s_{-i-j}) < u_j(s)$  if and only if  $(\delta - c) a^1 + (1 - \delta - c)(n^0 - 1) + (1 - \delta - c) < (1 - \delta + \lambda - c) a^1 + (\delta + \lambda - c)(n^0 - 1)$ , or equivalently,  $a^1 > \frac{1-2\delta-\lambda}{1-2\delta+\lambda} (n^0 - 1) + \frac{1-\delta-c}{1-2\delta+\lambda}$ . Note, however, that  $\frac{1-\delta-c}{1-2\delta+\lambda} \in (0, 1)$  and  $\frac{1-2\delta-\lambda}{1-\delta+\lambda-c} > \frac{1-2\delta-\lambda}{1-2\delta+\lambda}$  implies  $a^1 = \left\lceil \frac{1-2\delta-\lambda}{1-\delta+\lambda-c} (n^0 - 1) \right\rceil + 1 > \frac{1-2\delta-\lambda}{1-2\delta+\lambda} (n^0 - 1) + \frac{1-\delta-c}{1-2\delta+\lambda}$ , and hence condition (2) holds also for  $j$ . According to Lemma 6,  $s \in S^{PNE}$ .  $\square$

# Bibliography

- I. Alger and J.W. Weibull. Homo moralis – Preference evolution under incomplete information and assortative matching. *Econometrica*, 81:2269–2302, 2013.
- O. Baetz. Social activity and network formation. *Theoretical Economics*, 10:315–340, 2015.
- A. Benjamin, G. Chartrand, and P. Zhang. *The Fascinating World of Graph Theory*. Princeton University Press, Princeton, New Jersey, 2015.
- H. Bester and W. Güth. Is altruism evolutionary stable? *Journal of Economic Behavior and Organization*, 34:193–209, 1998.
- F. Bolle. Is altruism evolutionary stable? And envy and malevolence? Remarks on Bester and Güth. *Journal of Economic Behavior and Organization*, 42:131–133, 2000.
- J.A. Bondy and U.S.R. Murty. *Graph Theory with Applications*. The Macmillan Press, London, 1977.
- Y. Bramoullé. Anti-coordination and social interactions. *Games and Economic Behavior*, 58(1):30–49, 2007.
- Y. Bramoullé and R. Kranton. Games played on networks. In Y. Bramoullé, A. Galeotti, and B. Rogers, editors, *The Oxford Handbook of the Economics of Networks*, pages 83–112. Oxford University Press, New York, 2016.
- Y. Bramoullé, D. López-Pintado, S. Goyal, and F. Vega-Redondo. Network formation and anti-coordination games. *International Journal of Game Theory*, 33(1):1–19, 2004.
- A. Cabrales, A. Calvó-Armengol, and Y. Zenou. Social interactions and spillovers. *Games and Economic Behavior*, 72:339–360, 2011.
- A. Calvó-Armengol, E. Patacchini, and Y. Zenou. Peer effects and social networks in education. *Review of Economic Studies*, 76:1239–1267, 2009.

- G. Charness, F. Feri, M. Meléndez-Jiménez, and M. Sutter. Experimental games on networks: Underpinnings of behavior and equilibrium selection. *Econometrica*, 82(5):1615–1670, 2014.
- S. Currarini, M.O. Jackson, and P. Pin. An economic model of friendship: Homophily, minorities and segregation. *Econometrica*, 77:1003–1045, 2009.
- E. Dekel, J.C. Ely, and O. Yilankaya. Evolution of preferences. *Review of Economic Studies*, 74:685–704, 2007.
- R. Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer, Berlin, 5th edition, 2017.
- B. Dutta and S. Mutuswami. Stable networks. *Journal of Economic Theory*, 76:322–344, 1997.
- L. Ellwardt, P. Hernández, G. Martínez-Cánovas, and M. Muñoz Herrera. Conflict and segregation in networks: An experiment on the interplay between individual preferences and social influence. *Journal of Dynamics and Games*, 3:191–216, 2016.
- J.C. Ely and O. Yilankaya. Nash equilibrium and the evolution of preferences. *Journal of Economic Theory*, 97:255–272, 2001.
- R.H. Frank. *Passion within Reason*. Norton, New York, 1988.
- A. Galeotti, S. Goyal, M.O. Jackson, F. Vega-Redondo, and L. Yariv. Network games. *Review of Economic Studies*, 77:218–244, 2010.
- B. Golub and Y. Livne. Strategic random networks and tipping points in network formation. Unpublished Manuscript, MIT, 2011.
- S. Goyal. Networks in economics. A perspective on the literature. In Y. Bramoullé, A. Galeotti, and B. Rogers, editors, *The Oxford Handbook of the Economics of Networks*, pages 47–70. Oxford University Press, New York, 2016.
- S. Goyal and F. Vega-Redondo. Network formation and social coordination. *Games and Economic Behavior*, 50:178–207, 2005.
- S. Goyal and F. Vega-Redondo. Structural holes in social networks. *Journal of Economic Theory*, 137:460–492, 2007.
- S. Goyal, P. Hernández, G. Martínez-Cánovas, F. Moisan, M. Muñoz Herrera, and A. Sánchez. Integration and diversity. *Experimental Economics*, 24:387–413, 2021.

- W. Güth. An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *International Journal of Game Theory*, 24:323–344, 1995.
- W. Güth and S. Napel. Inequality aversion in a variety of games: An indirect evolutionary analysis. *The Economic Journal*, 116:1037–1056, 2006.
- W. Güth and M. Yaari. Explaining reciprocal behavior in simple strategic games: An evolutionary approach. In Witt U., editor, *Explaining Process and Change*, pages 22–34. University of Michigan Press, Ann Arbor, MI, 1992.
- Y. Heller and E. Mohlin. Coevolution of deception and preferences: Darwin and Nash meet Machiavelli. *Games and Economic Behavior*, 113:223–247, 2019.
- P. Hernández, M. Muñoz Herrera, and A. Sánchez. Heterogeneous network games: Conflicting preferences. *Games and Economic Behavior*, 79:56–66, 2013.
- P. Hernández, G. Martínez-Cánovas, M. Muñoz Herrera, and A. Sánchez. Equilibrium characterization of networks under conflicting preferences. *Economic Letters*, 155:154–156, 2017.
- F. Herold and C. Kuzmics. Evolutionary stability of discrimination under observability. *Games and Economic Behavior*, 67:542–551, 2009.
- T. Hiller. Peer effects in endogenous networks. *Games and Economic Behavior*, 105:349–367, 2017.
- J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998.
- M.O. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, 2008.
- M.O. Jackson and A. van den Nouweland. Strongly stable networks. *Games and Economic Behavior*, 51(2):420–444, 2005.
- M.O. Jackson and A. Watts. On the formation of interaction networks in social coordination games. *Games and Economic Behavior*, 41(2):265–291, 2002.
- M.O. Jackson and Y. Zenou. Games on networks. In P. Young and S. Zamir, editors, *Handbook of Game Theory*, volume 4. Elsevier Science, 2014.
- M. Kearns, M.L. Littman, and S. Singh. Graphical models for game theory. In J.S. Breese and D. Koller, editors, *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, 2001.

- L. Kockesen, E.A. Ok, and R. Sethi. Evolution of interdependent preferences in aggregative games. *Games and Economic Behavior*, 31:303–310, 2000.
- N.V.R. Mahadev and U.N. Peled. *Threshold Graphs and Related Topics*. North-Holland, Amsterdam, 1995.
- A. Mauleon and V. Vannetelbosch. Network formation games. In Y. Bramoullé, A. Galeotti, and B. Rogers, editors, *The Oxford Handbook of the Economics of Networks*, pages 167–190. Oxford University Press, New York, 2016.
- J. Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, 1982.
- J. Maynard Smith and D. Harper. *Animal Signals*. Oxford University Press, Oxford, 2007. Oxford Series in Ecology and Evolution.
- J. Maynard Smith and G.R. Price. The logic of animal conflict. *Nature*, 246:15–18, 1973.
- S. Morris. Contagion. *Review of Economic Studies*, 67:57–78, 2000.
- R. Myerson. Graphs and cooperation in games. *Mathematics of Operations Research*, 2: 225–229, 1977.
- J. Oechssler and F. Riedel. On the dynamic foundation of evolutionary stability in continuous models. *Journal of Economic Theory*, 107:223–252, 2002.
- E.A. Ok and F. Vega-Redondo. On the evolution of individualistic preferences: An incomplete information scenario. *Journal of Economic Theory*, 97:231–254, 2001.
- O. Orlova. Personal preferences in networks. Center for Mathematical Economics Working paper, 631. Bielefeld: Center for Mathematical Economics, 2019.
- E. Patacchini and Y. Zenou. Juvenile delinquency and conformism. *Journal of Law, Economics and Organization*, 28:1–31, 2012.
- N. Robalino and A.J. Robson. The economic approach to the "theory of mind". *Philosophical Transactions of the Royal Society B (Biological Sciences)*, 367:2224–2233, 2012.
- A.J. Robson. Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Theoretical Biology*, 144:379–396, 1990.
- A.J. Robson and L. Samuelson. The evolutionary foundations of preferences. In J. Benhabib, A. Bisin, and M.O. Jackson, editors, *The Social Economics Handbook*, pages 221–310. North Holland, 2011.

- G. Ruxton, T.N. Sherratt, and M.P. Speed. *Avoiding Attack: The Evolutionary Ecology of Crypsis, Warning Signals and Mimicry*. Oxford University Press, Oxford, 2004.
- L. Samuelson. Introduction to the evolution of preferences. *Journal of Economic Theory*, 97: 225–230, 2001.
- M.E. Schaffer. Evolutionary stable strategies for a finite population and a variable contest size. *Journal of Theoretical Biology*, 132:469–478, 1988.
- R. Sethi and E. Somanthan. Preference evolution and reciprocity. *Journal of Economic Theory*, 97:273–297, 2001.
- F. Vega-Redondo. Links and actions in interplay. In Y. Bramoullé, A. Galeotti, and B. Rogers, editors, *The Oxford Handbook of the Economics of Networks*, pages 191–214. Oxford University Press, New York, 2016.
- J.W. Weibull. *Evolutionary Game Theory*. MIT Press, Cambridge, MA, 1995.
- W. Wickler. *Mimicry in Plants and Animals*. McGraw-Hill, New York, 1968.
- T. Wiseman and O. Yilankaya. Cooperation, secret handshakes, and imitation in the prisoners’ dilemma. *Games and Economic Behavior*, 37:216–242, 2001.

