Social preferences on networks*

Sarah Rezaei† — *Utrecht University*

Stephanie Rosenkranz‡ — *Utrecht University*

Utz Weitzel§ — *VU University Amsterdam*

Bastian Westbrock¶ — *Hamburg University*

March 5, 2024

Abstract

Social preferences are a powerful determinant of human behavior. We study their behavioral implications within the context of a network game. A key feature of our game is the existence of multiple equilibria that widely differ in terms of their payoff distributions. Determining which equilibrium is most plausible is thus a key concern. We show that introducing social preferences into the game can resolve the problem of equilibrium multiplicity. However, the selected equilibria do not necessarily yield more efficient or egalitarian payoff distributions. Rather, they just reinforce the inequality that is already inherent in a network structure. We validate these predictions in an experiment and discuss their implications for managerial practice and behavior in larger networks.

JEL: D85, C70, C91, H41

Keywords: social preferences, network games, public goods

1 Introduction

In our daily lives, we are involved in many social interactions and constantly struggle to divide our time, effort, and resources with others. The time and effort we spend in this way can be viewed as a local public good that we share with our interaction partners. To give some examples, the preventive measures we take in a pandemic to

*This study belongs to the research program "Cooperation in Social and Economic Relations" at Utrecht University.

†Utrecht University School of Economics, t.rezaei@uu.nl

‡Utrecht University School of Economics, s.rosenkranz@uu.nl

§Free University of Amsterdam & Tinbergen Institute & Radboud University, Institute for Management Research, u.weitzel@vu.nl

¶Hamburg University, Department of Economics and Social Sciences, bastian.westbrock@uni-hamburg.de

protect our contacts, the time we dedicate to a joint project with our co-workers, or our experimentation with new technologies, which reduces the adoption costs for others, all these investments can be viewed as our contribution to a local public good.

Not all of us have access to the same interaction partners, however, and so, not all of us have access to the same public goods investments by others. The network structure of social interactions thus has major consequences for the distribution of the costs and benefits within a group or society. This is where social preferences come into play. Numerous experimental and empirical studies have consistently demonstrated that social preferences shape our behavior in the provision of public goods, especially within small groups. A recurring finding is that individuals contribute higher and more equitable amounts to these groups than what would be anticipated from a purely selfish viewpoint (see, e.g., Andreoni and Bernheim, 2009; Eckel and Harwell, 2015).As such, social preferences may indeed have the potential to also overcome the inequalities in our social interactions.

It is not clear, however, how social preferences play out in a network of interdependent public goods. We study this topic for the first time in both theory and experiment. Our starting point is the seminal public goods game by Bramoullé and Kranton (2007), which shares many similarities with the social dilemmas described above: Players are embedded in a fixed network and make investments in a local public good shared with their direct network neighbors. In particular, there exists a privately optimal level of the good that even a pure payoff maximizer would contribute to. The key question, therefore, is who is willing to provide the public good and who is going to free ride. To frame it in game theory terms, the game has multiple Nash equilibria that significantly differ in terms of total welfare and the payoff distributions they induce.

Three important questions emerge from here: Do social preferences help to maintain public good investments beyond the private optima? Do they resolve the problem of equilibrium multiplicity in this game? And, if so, do they facilitate more equitable or more efficient payoff distributions when a network structure itself is asymmetric? To structure our thoughts on these questions, we first extend the Bramoullé and Kranton

(2007) game by allowing the players to possess other-regarding preferences. Specifically, we adopt the utility model proposed by Charness and Rabin (2002), which encompasses various different social preference types that real people have been shown to care about, including altruism, inequity aversion, and competitiveness, among others.[1] We then study the Nash equilibria of our modified game.

Our main result is as follows: Many of the Nash equilibria that emerge in the original game with payoff-maximizing players are no longer sustainable when players possess social preferences. Specifically, the key insight is that when players' social preferences satisfy certain conditions explained below, their strive for a certain payoff ordering in their local network neighborhood leads to a significant simplification and sharpening of the equilibrium predictions. In a Nash equilibrium on a star network, for instance, an inequity-averse player in the center position must earn more than the players in the periphery positions when at least one of them is inequity-averse as well. In the original game, by contrast, a second equilibrium exists where the center player earns less than everybody else. Similarly, in any equilibrium on a fully connected network, a group of inequity-averse players will invest the same and consequentially earn the same, while in the original game, a wide range of investment profiles can be supported. The underlying mechanism in both examples is that the socially concerned players share a common understanding of which equilibrium to play.

There are two important aspects of this result that we would like to stress here. Firstly, many of our predictions are robust to the "strength" of players' social preferences, that is, the weights they assign to other players' payoffs. This is important because it means that our predictions can be applied to various social contexts, irrespective of whether social comparison concerns play a prominent or minor role there. Nevertheless, our predictions become actually even sharper when players have weaker social preferences. In fact, in the limit of marginal comparison concerns, our predicted equilibrium set is even a proper subset of the equilibrium set in the original game for

[1] See Bruhin et al. (2019) and Falk et al. (2018) for empirical evidence on the diversity of social preferences, and Kerschbamer and Müller (2020) and Reuben and Riedl (2013) for how this diversity can, for instance, explain differences in political attitudes or contribution norms.

many of the networks we study. In this sense, introducing social preferences into the Bramoullé and Kranton (2007) game results in equilibrium selection.

Secondly, however, we find that social preferences do not always lead to a refined equilibrium set across all networks and for all types of social preferences. Rather, this is tied to two conditions. First, players must have what we term compatible social preferences, which means that their preferences need to align with their positions in a network. Preference compatibility is satisfied, for instance, when all players in a network are competitive, inequity averse, or have social welfare concerns. It is violated, by contrast, when an altruistic player in the center position of a star interacts with a group of competitive players in the periphery positions. In such cases, the equilibrium set might even be larger than in the original game. Second, the network in which players are embedded must be nested in the sense that the neighborhoods of some players in the network must be contained in the neighborhoods of others (Mariani et al., 2019). In particular, the ideal constellation for our predictions to apply is when one player nests the neighborhoods of all other players as in the star. In contrast, our theory predicts no refined equilibrium set when no player nests the neighborhood of any other player, such as in the circle network. Here, players cannot agree on the equilibrium to be played, irrespective of their social preferences.

In the second part of our paper, we validate the key predictions and mechanisms behind our theory in an experiment. Our tests leverage one of the useful features of the two conditions behind our predictions, nestedness and preference compatibility, namely that they are readily measurable. Consequently, we compare investment profiles across networks with varying degrees of nestedness and among subject groups exhibiting different a-priori elicited social preference combinations to assess whether the observed investments move as predicted. Our experimental design incorporates two additional features to facilitate the test: first, a large strategy space allowing for the full set of Nash equilibria and deviations thereof to emerge and, second, a continuous-time framework, enabling subjects to freely adjust their choices over a specific time interval.

To provide an outlook on our findings, we do not observe any evidence suggesting that social preferences lead to a more equitable or more efficient payoff distribution than expected from a group of purely payoff-maximizing players. Instead, the majority of investments in our experiment closely align with the equilibrium predictions for the original game. Nevertheless, groups with compatible social preferences managed to coordinate their choices in two aspects better: they reached the predicted equilibrium profiles more frequently, and they converged to their final investments in a shorter time.

In the next section, we relate our contribution to the existing literature. Section 3 delves into our theoretical predictions, Section 4 outlines the experimental design, and Section 5 analyzes our findings. In Section 6, we explore the practical implications of our results for managerial practices and the broader social interaction networks that inspired our study. Section 7 concludes. The proofs of all our formal statements, supplementary evidence from the experiment, and the replication instructions can be found in the appendix.

2 Related literature

Our paper relates to the literature on social preferences and social networks. In the domain of social networks, our primary contribution lies in being the first to theoretically explore a network game with socially concerned players. While a few earlier theories have studied settings of socially concerned agents in a network, most notably Ghiglino and Goyal (2010), Immorlica et al. (2017), and Bourlès et al. (2017), a key distinction lies in their focus on contexts devoid of any strategic interactions between the agents, if it were not for their social comparison concerns. Their motivations stem from peer comparisons in otherwise anonymous markets, financial transfers between family members, or an individual's status in a large neighborhood. In contrast, we study a game where players not only observe but also influence each other, resulting in complex interactions and multiple strategy profiles in Nash equilibrium.[2] Thus,

[2]One other notable exception of a network game with socially concerned players is the paper by Richefort (2018). Nevertheless, similar to all the other theories, also his game yields a unique equilibrium

by resolving the issue of equilibrium multiplicity, social preferences play an entirely different role in our theory.

With this finding, we also contribute to another important branch in the theoretical networks literature that aims to tackle the pervasive problem of equilibrium multiplicity. As emphasized by Bramoullé et al. (2014) and Allouch (2015), this issue is most severe in games where players' actions are strategic substitutes, so precisely the class of games looked at in our study. Previous efforts to resolve the problem have considered Nash tâtonnement stability (Bramoullé and Kranton, 2007), stochastic stability (Boncinelli and Pin, 2012), and limited information about the network structure (Galeotti et al., 2010) as equilibrium refinement concepts. While their predictions broadly coincide with those derived from our theory for all the star-like networks, our theory provides additional insights into phenomena unexplained by previous theories. For instance, it is able to explain why individuals tend to split their investments equally when they interact in pairs or why they fail to coordinate their choices within loosely connected local interaction structures, such as the circle network. Both these phenomena, while empirically very relevant, remained previously unaddressed.[3]

In the experimental networks literature, the central question mirrors that of the theory: which equilibrium prevails on which network structure, and why? Yet, the empirical support for the aforementioned theories remains, at best, mixed. For instance, Charness et al. (2014) delved into the role of incomplete information about the network structure, finding that it does not inherently facilitate coordination. Instead, in their experiment, risk dominance emerged as the guiding principle for equilibrium selection. Moreover, in an experimental design similar to ours, Rosenkranz and Weitzel (2012) compared the predictions of Nash tâtonnement stability, risk dominance, and quantal-

point regardless of whether the players are socially concerned or not. As such, social preferences merely "shift" the unique equilibrium point in his theory, whereas in our theory, they play a crucial role in helping players decide which equilibrium to coordinate on. Another noteworthy distinction lies in the fact that all earlier theories, including Richefort's, focus on one specific type of social preference, such as altruism or competitiveness. We, in contrast, look at the empirically more relevant case of preference heterogeneity.

[3]For instance, equal sharing is the by far most common outcome in the two-player public goods games reviewed in Andreoni and Bernheim (2009). Moreover, the experiments of Berninghaus et al. (2002) and Cassar (2007) made clear how difficult it is to coordinate on loosely connected local interaction structures.

response theory, providing no more than partial support for all three concepts.[4]

Both of these experiments share a common limitation: social preferences have never been given a chance to reveal their full potential as an equilibrium-refinement device. One reason is that much of their evidence is derived from games on asymmetric networks, where all the existing refinement concepts, including ours, predict just the same equilibrium. Another issue arises from their use of a binary strategy space that precludes equal divisions by design or their implementation of a simultaneous choice format, making coordination difficult in the complex environment of a network game. In contrast, we follow Berninghaus et al. (2002) and Goyal et al. (2017) in implementing a continuous-time version of the Bramoullé and Kranton game. This version maintains the large strategy space of the original game while still enabling coordination, as players can learn about their investments before the final payout period.

Our paper is finally related to the extensive literature on social preferences. It is particularly close to an emerging group of studies that goes beyond the influence of social-comparison concerns in standard linear public goods or bargaining games. Similar to us, also these studies find a major role of social preferences in coordinating our choices. Binmore (2005), for example, argues that they help us navigate unfamiliar social dilemmas. Moreover, Reuben and Riedl (2013) and Fehr and Schurtenberger (2018) demonstrate how social preference influences the foundation of social norms, and Kahneman et al. (1986) and Eyster et al. (2021) illustrate their impact on a market's resistance to change. Closest to our study, Dufwenberg and Patel (2017) present a theoretical model showing how social preferences can reduce the number of Nash equilibria in a threshold-level public goods game. However, the arguments underlying their result differ entirely from ours. Moreover, while their theory speaks to public goods provision in small communities, the application we have in mind is the alloca-

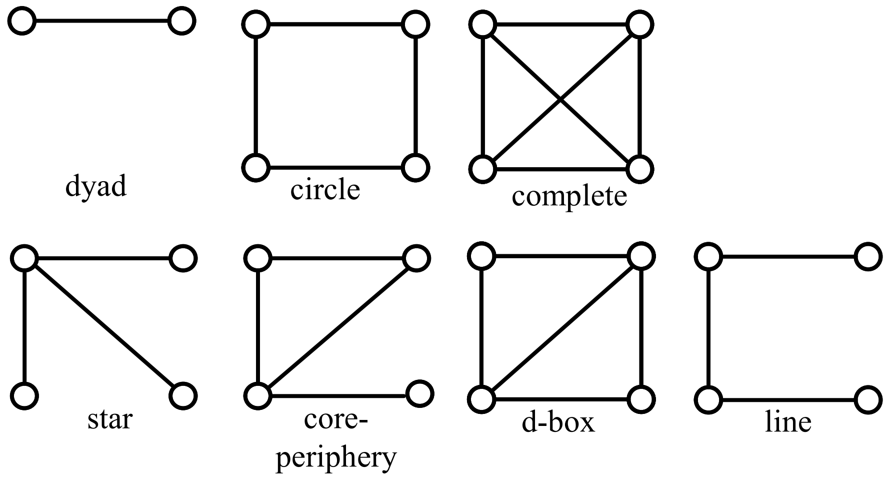
[4]The only other experimental study on the role of social preferences in networks that we are aware of is Zhang and He (2021). However, much like the theory papers mentioned above, they study a dominant-strategy game, where social preferences merely shift the observed investments. Social preferences in our context, by contrast, make up the difference between a center- or a periphery-specialized equilibrium and, thus, between being the sole contributor to a public good or a free rider. Moreover, we should mention another related line of experimental work investigating the influence of communication in network games (Choi and Lee, 2014; Charness et al., 2023). This work has revealed another effective means of coordination.

tion of scarce resources in a network of interdependent public goods.

3 Theory

3.1 The rules of the basic game

Figure 1: Networks in the experiment



We study the role of social preferences in the Bramoullé and Kranton (2007) local public goods game. The rules are as follows: n players are embedded in a fixed network g, some of which are illustrated in Figure 1. All players simultaneously select an investment $e_i \in [0, \bar{e}]$ that contributes to their own local public good and that of their direct neighbors in g.[5] Let $e_{-i} = (e_j)_{j \neq i}$ represent the investments of all players except player i, and let $N_i = \{j \in N \backslash \{i\} : ij \in g\}$ indicate the set of players in i's neighborhood. Player i's payoff is determined by the following expression:

$$\pi_i(e_i, e_{-i}) = b\big(e_i + \sum_{j \in N_i} e_j\big) - ce_i \,. \tag{1}$$

Here, $c > 0$ denotes the investment cost per unit and $b(\cdot)$ the social benefit function. This function is strictly increasing and concave on $[0, n\bar{e}]$ and satisfies $b(0) = 0$ and $b'(0) > c > b'(\bar{e})$. In most parts of our theory, we will more concretely assume that

[5] One might think of these partner-independent investments as the efforts in organizing parties for friends, the experimentation with new tools, or neighborhood beautification efforts, all vis-à-vis the time a person spends on her own personal projects.

$b(\cdot)$ is a quadratic function with $|b''| > (2b'(0) - c)/\bar{e}$, so that, regardless of a player's "strength" of social preferences, no player ever invests $\bar{e}$.

There are two important observations to be made about the Bramoullé and Kranton game. First, the exists a positive investment level e^* in the game, defined by $e^* = (b')^{-1}(c)$, that even a payoff-maximizing player would be willing to contribute to if the sum of her neighbors' investments is smaller. As a result, the investments of any two neighbors are strategic substitutes because the higher the neighboring investments, the less a player has to contribute herself to fill the gap until e^*.

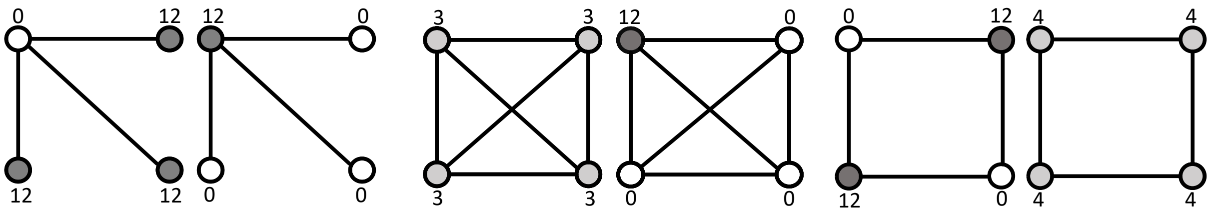
Second, every network structure has multiple Nash equilibria as illustrated in Figure 2 for three of the networks in our experiment (where $e^* = 12$). Most strikingly, the Nash equilibria differ markedly in terms of both the investment and payoff distributions they induce among players. For instance, in the star, core-periphery, line, and d-box of Figure 1, the equilibrium set includes both a *center-specialized* public good, where the center player invests $e_c = e^*$ and all the other players free ride, and a *periphery-specialized* public good, where the center player free rides on the investments of others. In the dyad and complete network, the equilibrium set even encompasses a continuum of profiles, ranging from an *equal-split* profile to a *specialized* equilibrium, where e^* is provided by a single player.

As this pattern emerges consistently on every other network structure as well, a major drawback of the game is that it does not predict any systematic relationship between the structure of a network and players' behavior within it. However, as we will see below, introducing social preferences into the game has the potential to significantly refine the equilibrium set. Moreover, the equilibria selected in this way are empirically highly relevant.

3.2 The social preference function

Social preferences are commonly understood as the human tendency to take the well-being of others into account when making a decision (e.g., Fehr and Schmidt, 1999). However, beneath this general tendency lies a significant amount of hetero-

Figure 2: Nash equilibria on three networks



NOTES: For the complete network, only two out of a continuum of Nash equilibria (with $\sum_{i \in N} e_i = e^* = 12$ as the only condition) are illustrated. In the circle network, there exists a third equilibrium where the players in the upper left and the lower right corners each invest 12.

geneity in terms of when and how individuals take other's well-being into account.[6]

To address this heterogeneity, the theoretical literature has developed various utility models aiming to capture social preferences within specific contexts (see Fehr and Charness, 2023, for a review). Our preferred model is an n-player extension of the distributional preference model by Charness and Rabin (2002) and Schulz and May (1989). This extension proves particularly useful as it nests many of the social preference types identified in the literature in a very parsimonious way.[7]

According to this model, a player's utility is as expressed as follows:

$$U_i(e_i, e_{-i}) \; = \; \pi_i \; + \; \frac{\sigma_i}{|R_i|} \sum_{j \in R_i^-} \left(\pi_j - \pi_i\right) \; + \; \frac{\rho_i}{|R_i|} \sum_{j \in R_i^+} \left(\pi_j - \pi_i\right). \tag{2}$$

Here, R_i denotes the player's reference group, and ρ_i and σ_i represent her social preference parameters, satisfying the conditions (i): $1 > \rho_i \geq \sigma_i > -1$ and (ii): $|\sigma_i| \geq |\rho_i|$ if $\rho_i > 0 > \sigma_i$.

A player's utility is thus a linear combination of her own material payoff π_i and a social preference component. The latter reflects the (dis-)utility a player derives from comparing her payoff with that of other players. With whom a player compares is defined by her reference group R_i. In our network context, it may be reasonable for

[6]See, for instance, the empirical evidence in Falk et al. (2018) and Bruhin et al. (2019).

[7]The distributional preference model also nests several of the utility functions used in the aforementioned literature on social networks. Notably, Ghiglino and Goyal (2010) assume what we term spiteful players, while Immorlica et al. (2017) assume what we refer to as competitive players. Bourlès et al. (2017), by contrast, develop a model wherein players know each other well, enabling them to incorporate each others' utilities, rather than just payoffs, into their own utility functions.

this group to just comprise the direct neighbors in a network (i.e., $R_i = N_i$), as these players can be directly influenced. Alternatively, it may be reasonable that a player also compares herself with players beyond her direct neighborhood, particularly in small networks. Our theory is flexible enough to accommodate both scenarios.

Regardless of the reference group's size, a player distinguishes between peers who are behind ($j \in R^+ = \{j \in R : \pi_j < \pi_i\}$) and peers who are ahead ($j \in R^- = \{j \in R : \pi_j > \pi_i\}$). The parameters ρ_i and σ_i then govern the (dis-)utility that a player derives from comparing her payoff with those behind and those ahead. In combination, these two parameters define various meaningful social preference types: Unconditional altruists ($\rho_i \geq \sigma_i > 0$), for instance, always assign a positive weight to their peers' payoffs, regardless of whether they are ahead or behind. Also, social welfare types ($\rho_i > \sigma_i = 0$) assign a positive weight to their peers' payoffs unless they earn less than everybody else in their reference group. In such a case, they behave like ordinary payoff maximizers, aiming to fill the gap between their neighbors' investments and e^*. In the negative domain, spiteful players ($0 > \rho_i \geq \sigma_i$) always assign a negative weight to their peers' payoffs. Competitive types ($0 = \rho_i > \sigma_i$), by contrast, behave like ordinary payoff maximizers when their payoffs are higher than everybody else's. The two domains are connected by the inequity-averse types ($\rho_i > 0 > \sigma_i$) who assign a positive or negative weight to their peers' payoffs depending on whether they are ahead or behind them. In sum, utility function (2) captures a broad spectrum of empirically relevant preference types and, as we will see below, it is also simple enough to generate sharp predictions within the context of a network game.

3.3 The rules of the modified game

Consistent with the broader empirical reality, and the specific context of our experiment in particular, we envision a game where players differ in their social preferences. Specifically, we assume that the social preference type of each player, denoted as $\tau_i = (\rho_i, \sigma_i, R_i)$, is determined before the start of the game through a random draw from a common support T, which we assume is a finite subset of the set of all types compatible with utility function (2). All players become aware of their own types, but they may only possess partial information about the types of the other players.

This naturally prompts the question of how much players know about the social preferences of others. Throughout the main text, we make the straightforward assumption that all players are completely informed about the preference types of every other player. As we will see in Section 5.1, this assumption, albeit highly stylized, yields predictions that are readily applicable in the context of our experimental game. Moreover, as demonstrated in Appendix A.2.5, our key results remain robust even in a richer setting where players have incomplete information about each other.

3.4 Equilibrium predictions

We are now ready to present the Nash equilibrium predictions for our modified game featuring socially concerned players, henceforth referred to as the other-regarding equilibria (ORE). A first notable observation is that our modified game readily lends itself to well-established fixed-point theorems, as summarized in Dasgupta and Maskin (1986). Consequentially, the game is guaranteed to have at least one pure-strategy ORE for any combination of player types $\tau = (\tau_1, .., \tau_n)$:

Proposition 1 (Equilibrium Existence). *The Bramoullé and Kranton game with socially concerned players and a quadratic social benefit function has at least one Nash equilibrium in pure strategies for every network g.*

Refer to Appendix A.1 for the proof. This raises several equally intriguing follow-up questions: How many OREs does the game possess on each network? How are the investments and payoffs distributed in these equilibria? And, crucially, how do both these aspects depend on the social preferences of the players? The following examples elucidate the fundamental intuition behind our main results.

Example 1 (Star with an Altruist in the Center). Let us first consider the 4-player star network depicted in Figure 1. Suppose that the center position is occupied by an unconditional altruist (with $\rho_c = \sigma_c > 0$), while the three players in the periphery are payoff maximizers (i.e., $\rho_p = \sigma_p = 0$). Then, one can readily show that the following three profiles describe all possible OREs:[8]

1. *center-specialized:* $e_c = \hat{e}(\rho_c)\,,\; e_p = 0$
2. *distributed:* $e_c = \frac{3e^* - e^*(\rho_c)}{2}\,,\; e_p = \frac{e^*(\rho_c) - e^*}{2}$ if $e^*(\rho_c) < 3e^*$
3. *periphery-specialized:* $e_c = 0\,,\; e_p = e^*$ if $e^*(\rho_c) < 3e^*$

Here, $e^*(\rho_c)$ denotes the total investment desired by the altruist when the payoff maximizers each make a positive investment, while $\hat{e}(\rho_c)$ depicts the altruist's total desired investment when the payoff maximizers do not invest, with $e^*(\rho_c) = (b')^{-1}(\frac{1-2\rho_c}{1-\rho_c}c) > \hat{e}(\rho_c) = (b')^{-1}((1-\rho_c)c) > e^*$.

This example illustrates one of the complicating factors introduced into the game when players are socially concerned. Not only can the original Nash equilibria of the Bramoullé and Kranton game be sustained as ORE, but additional equilibria may emerge that are not Nash when the players have pure monetary concerns. In Example 1, this is exemplified by the distributed equilibrium profile where all four players make a positive contribution to the public good. This is an ORE because the altruist is willing to maintain a total desired investment in her neighborhood that is greater than e^*. Hence, even if all the money maximizers invest $e_p = e^* - e_c > 0$, the altruist is still inclined to make an extra contribution. Only in the extreme case where the altruist cares a lot about the payoffs of the other players (i.e., when $e^*(\rho_c) \geq 3e^*$) does the ORE set collapse to a unique equilibrium where the altruist solely provides the public good. We do not pay much attention to this case, however, because it is unlikely that any individual is so altruistic.[9]

As demonstrated in our next example, social preferences indeed have the potential to narrow the equilibrium set, even under the more realistic condition of small to moderate social preferences.

[8]For the distributed ORE profile, use the first-order conditions $\frac{\partial U_c(e)}{\partial e_c} = \big(b'(e_c + 3e_p) - c\big)(1-\sigma_c) + \sigma_c b'(e_c + e_p) = 0$ and $\frac{\partial \pi_p(e)}{\partial e_p} = b'(e_c + e_p) - c = 0$.

[9]See, for instance, Figure 4 for evidence on this.

Example 2 (Star with a Spiteful Player in the Center). Let's revisit the star network again, with three payoff maximizers in the peripheral positions. However, this time, the center player is of a competitive or spiteful type (with $\sigma_c < 0$). In this case, the game has a unique *periphery-specialized* ORE, where $e_c = 0$ and $e_p = e^*$.

Why are the other two profiles of Example 1 no equilibria anymore when the star center player is of a competitive or spiteful type? Suppose the center player would make a positive contribution as in these profiles. His total desired investment would be no larger than e^* when he is competitive, and it would even be less than e^* when he is spiteful or when he invests the lion's share so that $\pi_c(e) < \pi_p(e)$. Hence, the periphery players must make a contribution themselves to fill the gap until their desired e^*. Yet, for any contribution in the periphery, the center player reduces his own investment, leading to further investment increases in the periphery, etc. Hence, a competitive or spiteful player in the center position of a star destabilizes any center-specialized or distributed profile.

Nevertheless, as our following example illustrates, this does not mean that competitive or spiteful players free ride in all network positions alike.

Example 3 (Circle with Spiteful Players). Consider a circle network with players labeled $1-4$. Suppose that players 1 and 3 are of a spiteful type (with $\rho_i = \sigma_i < 0$) and suppose they compare their payoffs only with their direct neighbors 2 and 4, who are again payoff maximizers. The set of ORE is in this scenario defined by

1. *1-and-3-specialized:* $e_1 = e_3 = \hat{e}(\sigma)\,,\ e_2 = e_4 = 0 \quad \text{if } 2\hat{e}(\sigma) > e^*$
2. *distributed:* $e_1 = e_3 = \frac{2e^* - e^*(\sigma)}{3}\,,\ e_2 = e_4 = \frac{2e^*(\sigma) - e^*}{3} \quad \text{if } 2e^*(\sigma) > e^*$
3. *2-and-4-specialized:* $e_1 = e_3 = 0\,,\ e_2 = e_4 = e^*$

where $e^*(\sigma)$ denotes the total desired investment of the spiteful players when the payoff maximizers make a positive contribution, and $\hat{e}(\sigma)$ their total desired investment when the payoff maximizers refrain from investing, with $e^*(\sigma) < \hat{e}(\sigma) < e^*$.

In other words, as long as players 1 and 3 are not too spiteful (i.e., $2\hat{e}(\sigma) > e^*$), the same large equilibrium set emerges on the circle as in the original game with payoff-maximizing players. In particular, there is an ORE where the public good is entirely provided by the spiteful players 1 and 3.

Our final example makes clear that the same large equilibrium set emerges on the circle under other preference constellations as well:

Example 4 (Circle with Inequity Averse Players). Consider the circle network again, but with inequity averse players (with $\rho_i > 0 > \sigma_i$ and $|\sigma_i| = |\rho_i|$) in all four positions. The ORE set is then given by[10]

1. *1-and-3-specialized:* $e_1 = e_3 = \hat{e}(\sigma)\,,\, e_2 = e_4 = 0 \quad \text{if } 2\hat{e}(\sigma) > \hat{e}(\rho)$
2. *distributed:* $\frac{e^*(\sigma)}{3} \leq e_i = e_j \leq \frac{e^*(\rho)}{3}$
3. *2-and-4-specialized:* $e_1 = e_3 = 0\,,\, e_2 = e_4 = \hat{e}(\sigma) \quad \text{if } 2\hat{e}(\sigma) > \hat{e}(\rho)$

So, why can the equilibria from the original game be supported as ORE on the circle regardless of the players' preference types, while only one of them survives on the star when a spiteful player occupies the center position? The answer lies in the distinct network structures. In the circle, each player's neighbors have one neighbor of their own which they do not need to share with the player. As a consequence, in the 1-and-3-specialized equilibrium, players 2 and 4 can access the investments of the spiteful players 1 and 3, while the latter only have access to their own investments. And as the total investment received by players 2 and 4 is beyond e^*, they are unwilling to make the extra contribution that would make players 1 and 3 reduce theirs. As a result, a public good sponsored by players 1 and 3 can be supported in an ORE on the circle.

[10] Note that there is no other distributed equilibrium profile besides the equal-split equilibrium. In particular, there is no ORE with $e_1 = e_3 > e_2 = e_4$ because the necessary first-order conditions,

$$\begin{aligned} \frac{\partial U_1}{\partial e_1}(e) &= \big(b'(e_1 + 2e_2) - c\big)\big(1 - \sigma\big) + \sigma b'(e_2 + 2e_1) \;=\; 0 \\ \frac{\partial U_2}{\partial e_2}(e) &= \big(b'(e_2 + 2e_1) - c\big)\big(1 - \rho\big) + \rho b'(e_1 + 2e_2) \;=\; 0\,, \end{aligned}$$

can only be satisfied simultaneously when $e_1 = e_2$.

The situation is different for the peripheral players in the star network who do not receive any investment that the spiteful player in the center position would not have access to as well. The crucial distinction from the circle network is that the center player in the star nests the neighborhoods of all other players in the following sense:

Definition 1 (Nestedness)**.** *Player i nests the neighborhood of player j when $N_j \cup \{j\} \subseteq N_i \cup \{i\}$.*

However, the stark difference in predictions between Examples 1 and 2 made clear that another additional condition must be satisfied for social preferences to lead to a refined equilibrium set. While the center player's spitefulness in Example 2 helped to refine the equilibrium set, the center's altruism in Example 1 did not. As previously explained, the key difference is that the spiteful player is determined to undo any payoff differences in his own disadvantage if there is a need to, while the altruist is not. More generally, equilibrium selection through social preferences requires that the more powerful nesting positions of a network should be occupied by competitive or spiteful types. Conversely, the weaker nested positions should be occupied by social-welfare or altruistic types because these types are willing to undo any payoff disadvantages for their more powerful neighbors (even though Example 2 demonstrated that payoff maximizers suffice as well). Inequity-averse types, finally, fit into any network position because they are willing to rectify both their own and their neighbors' payoff disadvantages. To summarize, the following combination of preference types is sufficient for social preferences to lead to a refined ORE set:

Definition 2 (Preference compatibility)**.** *Consider two neighbors i and j in a network such that i nests the neighborhood of j. We say that their preferences are compatible with their network positions if $\tau_i \in T_c$ and $\tau_j \in T_p$, where*

$$\left(T_c = \{\textit{inequity averse, competitive, spite}\} \quad AND \quad T_p = T \backslash \{\textit{spite}\} \right)$$
$$OR$$
$$\left(T_c = T \backslash \{\textit{altruist}\} \quad AND \quad T_p = \{\textit{altruist, social welfare, inequity averse}\} \right).$$

Finally, while not crucial for most of our predictions, they can oftentimes be significantly refined when all players in the game possess small social preferences (i.e., their ρ_i- and σ_i-parameters are small in absolute terms). As demonstrated by our previous examples, the Nash equilibria of the original game and the corresponding ORE of the same type typically diverge in their precise investment levels, and these differences can grow substantial when players possess strong social preferences. Determining the complete ORE set can, therefore, be intricate in these cases. Therefore, it is often more fruitful to first confine the ORE set based on the maximum deviation that each player is willing to make from a best-response investment of a payoff maximizer. We refer to this as a player's social preference strength.

To formally define it, let $f_i(\tau_i, e_{-i})$ denote the best-response investment of a type-τ_i player in network position i, and let $f_i(e_{-i})$ denote the best response of a payoff maximizer in the same position. We say that

Definition 3 (Social preference strength)**.** *The social preference strength of a type-τ_i player in network position i is given by the smallest $\epsilon_i \in \mathbb{R}_+$ to satisfy*

$$\epsilon_i \;=\; \max\left\{ \left| f_i(\tau_i, e_{-i}) - f_i(e_{-i}) \right| \;:\; \forall\, e_{-i} \in [0,\, \bar{e}\,]^{n-1} \right\}.$$

The social preference strength of all players in a network is then given by $\epsilon \equiv \max_{i \in N}\{\epsilon_i\}$.

In the following, we will utilize our three definitions to fully characterize the ORE sets for all seven networks in our experiment. The proofs of our statements are provided in Appendix A.2.

3.4.1 Star, core-periphery, and d-box

The star, core-periphery, and d-box are the three networks in our experiment where one or two players nest the neighborhoods of all the other players. Yet, when all players are payoff maximizers, this fact has little impact on the structure of equilibria, as both periphery- and center-sponsored public goods can emerge in equilibrium, where the center player(s) earn(s) the most or the least, respectively.

The same ambiguous predictions emerge from our theory when the players in these networks are socially concerned but have the wrong preference combination. In contrast, the set of ORE is much refined when the players hold social preferences that are compatible with their respective network positions. Specifically,

- in the star: $\tau_c \in T_c$ for the center player and $\tau_p \in T_p$ for at least one peripheral player p,
- in the core-periphery: $\tau_c \in T_c$ for the center player and $\tau_j \in T_p \backslash \{inequity\,averse,$ $competitive\}$ for at least one non-center player j, and
- in the d-box: $\tau_c \in T_c \backslash \{inequity\,averse,\ social\,welfare\}$ for both centers $c \in C$ and $\tau_p \in T_p \backslash \{inequity\ averse,\ competitive\}$ for at least one peripheral player p.

In this case, *no* center-specialized equilibrium (with $e_j = 0$ for all $j \in N \backslash C$) can emerge in an ORE because the center player(s) must earn more than at least one other player:

$$\pi_c(e) \quad \geq \quad \min_{j \in N \backslash C} \{\pi_j(e)\} \quad \text{for all } c \in C\,. \tag{3}$$

The intuition extends immediately from Example 2. The ORE set can be refined even further on the star, core-periphery, or d-box when all four players possess small social preference concerns. The intuition is simple as well. Condition (3) cannot be satisfied in any distributed investment profile, where all four players make a positive contribution, given that ϵ is smaller than some critical value $\bar{\epsilon}$, with $\bar{\epsilon}^{dbox} < \bar{\epsilon}^{star} = \bar{\epsilon}^{core}$ defined in Appendix A.2. In other words, when players' social preferences are sufficiently small, an ORE must entail a periphery-specialized ORE, where the public good is entirely sponsored by the non-center players:

$$\textit{periphery-specialized:} \quad e_c = 0\,, \quad e_p \in [e^* \pm \epsilon]\,, \quad \text{and} \sum_{d \in D} e_d \in [e^* \pm \epsilon]\,. \tag{4}$$

In the limit of $\epsilon \to 0$, the ORE set, thus, even becomes a proper subset of the Nash equilibria of the original game.

3.4.2 Line

The two center players in the line network each have one periphery player whose neighborhood they nest. When all four players are payoff maximizers, this has, just as in the star, core-periphery, and d-box, little impact on the structure of equilibria because the only requirement on a Nash equilibrium is that $e_{p_i} = e^*$, $e_{c_i} = 0$, and $e_{c_j} + e_{p_j} = e^*$ for $i \in \{1, 2\}$ and $j \neq i$.

Again, the same holds true when players are socially concerned, but have the wrong preference combination. Also, in this case, an ORE might entail a periphery-specialized (with $e_{p_j} \geq e_{c_j}$) or a distributed (with $e_{p_j} < e_{c_j}$) investment profile. However, when players' social preferences are compatible with their respective line positions, and they possess sufficiently weak social preference, that is, when $\tau_c \in T_c$ for both line middle players, $\tau_p \in T_p$ for both end players, and $\epsilon < \bar{\epsilon}^{line} = e^*/5$, then an ORE must be

$$\textit{periphery-specialized:} \quad \pi_{c_i}(e) \geq \pi_{p_i}(e) \quad \text{and } e_{p_i} \geq e_{c_i} \quad \text{for } i \in \{1, 2\}. \tag{5}$$

Hence, social preferences can also resolve the problem of equilibrium multiplicity on the line network. Yet, they do so less effectively than on the star, core-periphery, or d-box because the fact that each center player only nests one other player's neighborhood means that all four players need to have compatible preferences for our equilibrium selection argument to apply.

3.4.3 Dyad and complete network

On the dyad and complete network, a wide range of investment profiles can be supported in a Nash equilibrium when players are payoff maximizers. The only requirement is that $\sum_{i \in N} e_i = e^*$.

Social preferences lead, in the first instance, to even more equilibria, as any profile can be supported in an ORE with arbitrary preference types as long as $\sum_{i \in N} e_i \in [e^* \pm \epsilon]$. However, our theory predicts a much more refined ORE set when each player meets both compatibility conditions of Definition 2. Specifically,[11]

- in the dyad: $\tau_i \in T_c \cap T_p$ for both $i = \{1, 2\}$,
- in the complete network: $\tau_i \in T_c \cap T_p$ for all $i \in N$ and all ρ_i-parameters, as well as all σ_i-parameters, are sufficiently close together.

In this case, our theory even predicts a unique ORE where the players split their total investment equally

$$\textit{equal-split:} \quad e_i = e_j = e \,, \quad \text{with } e \in [\frac{e^* \pm \epsilon}{n}] \,. \tag{6}$$

The intuition is straightforward. Suppose that not all investments are equal. The fact that players' neighborhoods are mutually nested means that the player with the highest investment earns weakly less than everybody else, while the player with the lowest investment earns weakly more. At least one of them would thus feel insulted in her understanding of fairness and, accordingly, adjust her investment up- or downward. Such adjustments can only be avoided when all players invest exactly the same.[12]

3.4.4 Circle

As already highlighted in Examples 3 and 4, the absence of nested neighborhoods in the circle puts an end to the equilibrium selection property of social preferences. All that can be said about the ORE set is summarized in these examples: It is as large as the equilibrium set of the original game, and it collapses with it when players' social preferences become small ($\epsilon \to 0$). A refined ORE set is just possible when players

[11]The condition $\tau_i \in T_c \cap T_p$ implies, for the dyad and the complete network, that (i) no player should be an altruist or a spiteful type, (ii) no two or more players should be payoff maximizers, and (iii) no two or more players should have distinct types from the set {payoff maximizer, social welfare, competitive}.

[12]Note that the strength of social preferences does not influence the emergence of an equal-split equilibrium on the dyad or complete network. It solely affects the extent by which the total investment of all players differs from e^*. A total investment of $ne > e^*$ can, for instance, be maintained by the aversion to guilt. As long as the material benefits from a downward deviation are smaller than the moral cost of guilt, and thus as long as ne is not too far away from e^*, players prefer their equilibrium investment e.

Note also that this equal-split prediction does not derive from any of the other established equilibrium refinement concepts, such as Nash tâtonnement stability, efficiency, or stochastic stability.

have certain combinations of strong social preferences, for instance, when two spiteful types interact with two money maximizers.

3.4.5 General networks

The insights gained from the previous analysis can be extended to arbitrarily large network structures, provided that the network has some nested neighborhoods and the players within these neighborhoods possess compatible social preferences. In this case, the following result applies:

Proposition 2. *Consider two players i and j in a nested neighborhood of a network g who have compatible social preferences, that is, $\tau_i \in T_c$ and $\tau_j \in T_p$. Then, in an ORE, player i (j) must earn weakly more (less) than at least one other player in i's (j's) neighborhood:*

$$\pi_i(e) \geq \min_{k \in N_i} \{\pi_k(e)\} \qquad OR \qquad \pi_j(e) \leq \max_{l \in N_j} \{\pi_l(e)\} \,. \tag{7}$$

While the result establishes rather weak bounds on the relative payoffs within a local network neighborhood, we already know how to strengthen it under some additional conditions on the network structure. For instance, when player j is solely connected to player i such as in the periphery position of a star, then $\pi_i(e) \geq \min_{k \in N_i}\{\pi_k(e)\}$. And, when i is also exclusively connected to j, then it even follows that $\pi_i(e) = \pi_j(e)$.

3.4.6 Network ranking

So far, we have seen that incorporating social preferences into the Bramoullé and Kranton (2007) game allows one to refine the equilibrium predictions for most of the networks in Figure 1. Crucially, we successfully eliminated all those equilibria that go against the intuitive expectations regarding the ranking of investments and payoffs in these networks, namely center-sponsored public goods in the star-like networks and unequal contributions in the dyad and complete network.

Our theory predicts more, however, because it also suggests systematic differences between these networks in terms of how likely a refined ORE can be expected to emerge on them when players are randomly assigned to their network positions, as in our experiment. One immediate implication of random assignment is that we would expact a large share of groups with compatible social preferences on the dyad compared to the complete network. The rationale is straightforward: in the dyad, it is easier to assemble a sufficient number of players who share a common understanding of fairness and, thus, a common understanding of which equilibrium to play. When we now add the plausible assumption that players coordinate on a random profile from the set of all ORE profiles consistent with their social preference types, $\tau = (\tau_1, ..., \tau_n)$, we arrive at our first prediction: the likelihood of observing an equal-split ORE is higher on the dyad than on the complete network,

$$P\big(\textit{equal-split} \,|\, g^{dyad}\big) \;\geq\; P\big(\textit{equal-split} \,|\, g^{comp}\big)\,. \tag{8}$$

Our theory also predicts marked differences between all the other networks. We already know from our analysis of the circle that in the absence of any nested neighborhoods, a network is prone to multiple equilibria. Thus, at least some degree of nestedness is a prerequisite for a refined ORE set. But even among the nested networks of Figure 1, there are some important differences. In particular, there is some asymmetry with regard to the ideal number of central positions (n^c) in a network, which nest other positions' neighborhoods, and the ideal number of peripheral positions (n^p), whose neighborhoods are nested. The larger n^c (e.g., comparing the star and the d-box), the more likely it is that an incompatible altruist or social-welfare type is assigned to one of the center positions, thus a type who is willing to provide the public good on her own. The larger n^c, therefore, the smaller the likelihood that we observe a periphery-specialized ORE in a network. The number of peripheral positions has the opposite effect. The larger n^p, the more likely it is that at least one altruist or social-welfare type is assigned to such a position, so a type who is willing to contribute to the public good if the central players invest less than e^*. The larger n^p, therefore, the higher the likelihood of a periphery-specialized equilibrium in a network.

Applied to our networks, we thus arrive at the following rankings:[13]

$$
\begin{aligned}
(i) \quad & P\big(\textit{periphery-spec.} \,|\, g^{star}\big) \;\geq \\
& \qquad \max\left\{P\big(\textit{periphery-spec.} \,|\, g^{dbox}\big)\,;\, P\big(\textit{periphery-spec.} \,|\, g^{line}\big)\right\}, \qquad (9)\\
(ii) \quad & P\big(\textit{periphery-spec.}|g^{core}\big) \;\geq\; P\big(\textit{periphery-spec.}|g^{dbox}\big)\,.
\end{aligned}
$$

Moreover, a refined ORE set is easier achieved in an asymmetric than in a symmetric nested network because, in the latter, players must match the preference compatibility requirements for both the nesting as well as the nested positions. We would, therefore, expect that[14]

$$
P\big(\textit{periphery-spec.} \,|\, g^{dbox}\big) \;\;\geq\;\; P\big(\textit{equal-split} \,|\, g^{comp}\big)\,. \qquad (10)
$$

Altogether, we thus arrive at the following testable predictions:

Hypothesis 1: In the networks of Figure 1, except the circle, a group of players with compatible social preferences is more likely to coordinate on a refined ORE, i.e., a profile satisfying (3)–(6), *than a group without compatible preferences.*

Hypothesis 2: Suppose that players are randomly assigned to network positions from a common pool of players. Then, the likelihood of observing a refined ORE on the seven networks of Figure 1 can be ranked according to the inequalities in (8)–(10).

Finally, for the circle network, our theory predicts that even if all the preference requirements of Definition 2 are met by a player group, this group does nevertheless not coordinate more likely on either a specialized or a distributed profile on the circle

[13] Beyond the intuition provided in the text, the rankings can be readily derived from the compatibility requirements outlined in Sections 3.4.1 and 3.4.2. These conditions also make clear why the line cannot be unambiguously compared to neither the core-periphery nor the d-box.

One can furthermore not rank the star and the core-periphery network because even though the compatibility requirements are stronger in the latter, the likelihood that a group with incompatible preferences hits a refined ORE profile by chance is higher on the core-periphery, as there are just more of these profiles.

[14] There is no comparable ordering of the line and the complete network because, for a compatible preference combination in the line, one requires that $\epsilon < \overline{\epsilon}^{line}$, while there is no such restriction on the social preference strength in the complete network.

than a group that does not match the criteria.

4 Experiment

We tested our hypotheses within a dynamic extension of the Bramoullé and Kranton (2007) game. Our choice is motivated by the insights gained from prior experiments on this game, which made it clear that many subjects find it difficult to coordinate their choices in any meaningful manner, especially in experiments that adopted the original large strategy space (e.g., Rosenkranz and Weitzel, 2012). As a substantive share of equilibrium play is essential for our theory testing, however, we implemented a continuous-time version of the game.

In particular, following the approaches of Callander and Plott (2005) and Berninghaus et al. (2006), every round of our game lasted between 30 and 90 seconds. During that time, subjects could continuously adjust their choices, choosing from the full set of positive integer values. Moreover, subjects received full information about the momentary investments and payoffs of every other player, which were updated five times per second (see Appendix C.2 for a screenshot).

Nevertheless, to adhere to the static environment of our theory, the actual payoffs in a round were solely determined by the momentary investments at the round ends. These ends were randomly determined by a draw from the uniform distribution on $[30, 90]$. At that point in time, investments were frozen and points were calculated based on the following linear-quadratic payoff function:

$$\pi_i(e) \;=\; \begin{cases} \left(e_i + \sum_{j \in N_i} e_j\right)\left(29 - e_i - \sum_{j \in N_i} e_j\right) - 5e_i & \text{if } e_i + \sum_{j \in N_i} e_j \leq 14 \\ 196 \;+\; e_i + \sum_{j \in N_i} e_j - 5e_i & \text{otherwise} \end{cases} . \quad (11)$$

As we will see below, equilibrium play was greatly facilitated by these design choices. A major factor certainly is that subjects did not need to formulate beliefs about the payoffs and investments of the other player because they could observe them directly.[15] At the same time, our implemented random stopping rule eliminated last-round effects.

4.1 Experimental procedure

We administered our experiment at the Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University, the Netherlands, in June 2008. The experiment was programmed in z-tree 3.0 (Fischbacher, 2007) and subjects were recruited via ORSEE (Greiner, 2015). A total of 120 students participated in eight sessions. No subject attended more than once.[16]

In a typical session, participants played each one of the seven networks illustrated in Figure 1 in one trial round and four payoff-relevant rounds. This resulted in a total of 960 network-level observations that we could use for our analysis: 120 rounds per four-player network (120 subjects divided by 4 players times 4 payoff-relevant rounds) and 240 rounds from the dyad. Each participant engaged in 28 payoff-relevant rounds, spent approximately 80 minutes in our laboratory, and earned, on average, 11.82 Euros, including a 3 Euro show-up fee.

Our choice for such a within-subject design was not only motivated by experimental efficiency but also because we could use our network game in this way to estimate our subjects' social preference parameters from it (see below for details). To mitigate the confounding impact of factors associated with our choice, we implemented several additional measures. Firstly, to minimize the impact of between-treatment spillovers, we adopted a balanced treatment order, ensuring that each network appeared equally often at different points in our sessions (see Table 11 in Appendix C.1 for the orders). Furthermore, to mitigate repeated game effects that typically emerge when the same groups of players interact multiple times (Andreoni, 1995; Fehr and Gächter, 2000),

[15]From a theoretical viewpoint, the observation of other players' investments and payoffs is, in fact, all a socially concerned player needs to know to formulate her own best-response investment. This is because utility function (2) is solely affected by the investments but not the preference parameters of the other players.

[16]Clearance for this experiment has been granted by the Ethical Review Committee of Utrecht University's Faculty of Law, Economics, and Governance. Further experimental details can be found in Appendix C.

we randomly reassigned our subjects to new groups and new network positions after every round.

4.2 Social preference elicitation

Key to our testing of Hypothesis 1 is that we also have an estimate of the social preference parameters of our subjects at hand. We estimated these parameters directly from their behavior in our network game.[17] Concretely, we assumed that in each round r and at each time point t , a subject chose an investment level $x \in \mathbb{N}_+$ to maximize

$$U_i(x, e_{-i,t-1,r}) \;+\; \theta_{i,x,t,r}\,, \tag{12}$$

where $U_i(\cdot)$ is the utility function in (2) and $\theta_{i,x,t,r}$ an iid type-1 extreme value distributed random utility component. A subject thereby only compared her payoff with that of her direct network neighbors, i.e., $R_{i,r} = N_{i,r}$.

Consequentially, we estimated, for each subject, the $(\hat{\rho}_i, \hat{\sigma}_i)$-pair that maximized the conditional likelihood for their actual sequence of investments (e_{it}) to be favored over any alternative sequence. For our estimations, we used all the available information from our experiment and, accordingly, estimated a subject's parameters based on her choices during all decision moments $t \in [30, t^{max}]$ in all the 28 payoff-relevant rounds in our experiment. For practical reasons, we limited the alternative investments to $x \in \{0, 1, 2, ..., 15\}$, however.[18]

With our estimated $(\hat{\rho}_i, \hat{\sigma}_i)$-pairs at hand, we then categorized our subjects based

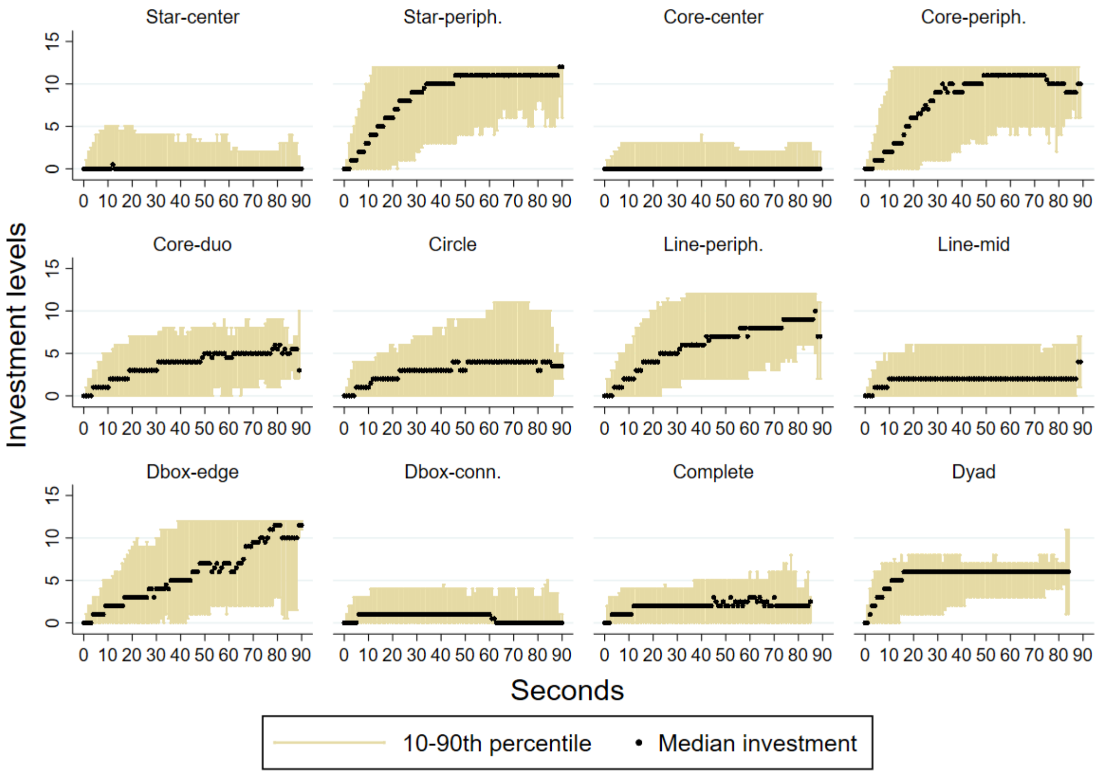
[17] There is mainly a practical reason for this. Our experiment was already 80 minutes long and we worried that subjects' fatigue would jeopardize the quality of our data collection if we added additional preference elicitation tasks. In fact, achieving the necessary precision in the parameter estimates would demand a considerable amount of time for these additional tasks. Bruhin et al. (2019), for instance, estimate a comparable utility model from 39 dictator games, a process that consumed at least 20 additional minutes in their experiment.

[18] This constraint is anyhow satisfied by 99.9% of all investments. Moreover, we ensured that the estimated $(\hat{\rho}_i, \hat{\sigma}_i)$-pair falls into the feasible range $1 > \rho_i \geq \sigma_i > -1$. Accordingly, we replaced the parameters of (2) by inverse logistic transformations of some deeper, unconstrained parameters $\rho_i^r, \sigma_i^r \in \mathbb{R}$:

$$\rho_i = -1 + 2\frac{\exp(\rho_i^r)}{1 + \exp(\rho_i^r)} \quad \text{and} \quad \sigma_i = -1 + 2\frac{\exp(\sigma_i^r)}{1 + \exp(\sigma_i^r)}\,,$$

and then solved our maximum likelihood function for $\hat{\rho}_i^r$ and $\hat{\sigma}_i^r$, computing heteroskedasticity-consistent standard errors.

Figure 3: Investments by network position over time



on their revealed social preference types and revealed preference strengths. For the preference type classification, we simply applied the parameter cutoffs presented in Section 3.2. For the preference strength classification, in turn, we utilized a theoretical result developed in Appendix A.3, which shows how to map a $(\hat{\rho}_i, \hat{\sigma}_i)$-pair into an upper bound $\hat{\epsilon}_i$ for a subject's true strength ϵ_i.[19]

5 Results

We first provide an overview of our experimental findings before we turn to our hypothesis tests.

5.1 Descriptive findings

We begin with a brief assessment of whether our static theory predictions make sense in the context of our dynamic experimental game. To this end, we plot in Figure 3 the evolution of the median investments and the 10–90 percentile investment ranges for each position in our seven networks over time. Clearly, with the exception of maybe the d-box edge position, the median investments in all positions converged consistently over time to some steady-state value, typically reached within the first 30 seconds already.[20] Furthermore, with the exception of possibly the positions in the circle network, the 10–90 percentile ranges shrank consistently over time, with an investment at the 90^{th}-percentile that never surpasses the total desired investment of $e^* = 12$ that maximizes our experimental payoff function (11). We see all this as confirmation that our subjects were myopically updating their choices in an attempt to reach an individually optimal investment in the payoff-relevant decision moments after 30 seconds. Accordingly, we interpret the evolution of investments as a best-response dynamic converging to a static equilibrium.

In support of this view, Figure 4 takes a closer look at the distributions of investments at the random round ends. The findings are much in line with our static ORE predictions. [21] For instance, the unique distributional modes in the two-player dyad and the complete network are, with three and six units respectively, consistent with the predicted equal-split equilibrium. Moreover, the prevalent zero contributions in the central positions of the star, core-periphery, d-box, and line, coupled with the substantial investments made in the peripheral positions of these networks, lend support to our anticipated periphery-specialized equilibrium. Even the somewhat dispersed pattern in the circle network, marked by minor peaks at zero, three, and twelve units,

[19]Obviously, we made several choices during our preference elicitation procedure, each carrying potential consequences for the precision of our parameter estimates. We discuss their implications for the purpose of our study are discussed at the end of Section 5.2.

[20]The seeming disturbance in this pattern after the 70-second mark, which is most pronounced in the d-box edge position, is simply due to the fact that many rounds ended before that time.

[21]The patterns in Figure 4 are highly robust. Very similar pictures emerge, for instance, when looking at the investment distributions across all payoff-relevant decision moments, $t \in [30, t^{max}]$, or when examining the investment distributions in the first and second halves of our experimental sessions separately. This reinforces our view that the findings in Figure 4 are not just driven by last-round effects or the specific order of networks in our sessions.

seems to be in line with our theoretical predictions, as we argued that social preferences would not help to solve the coordination problem on this network. Thus, a first glance at the data suggests that the behavior in our experiment is much in line with our static ORE predictions.

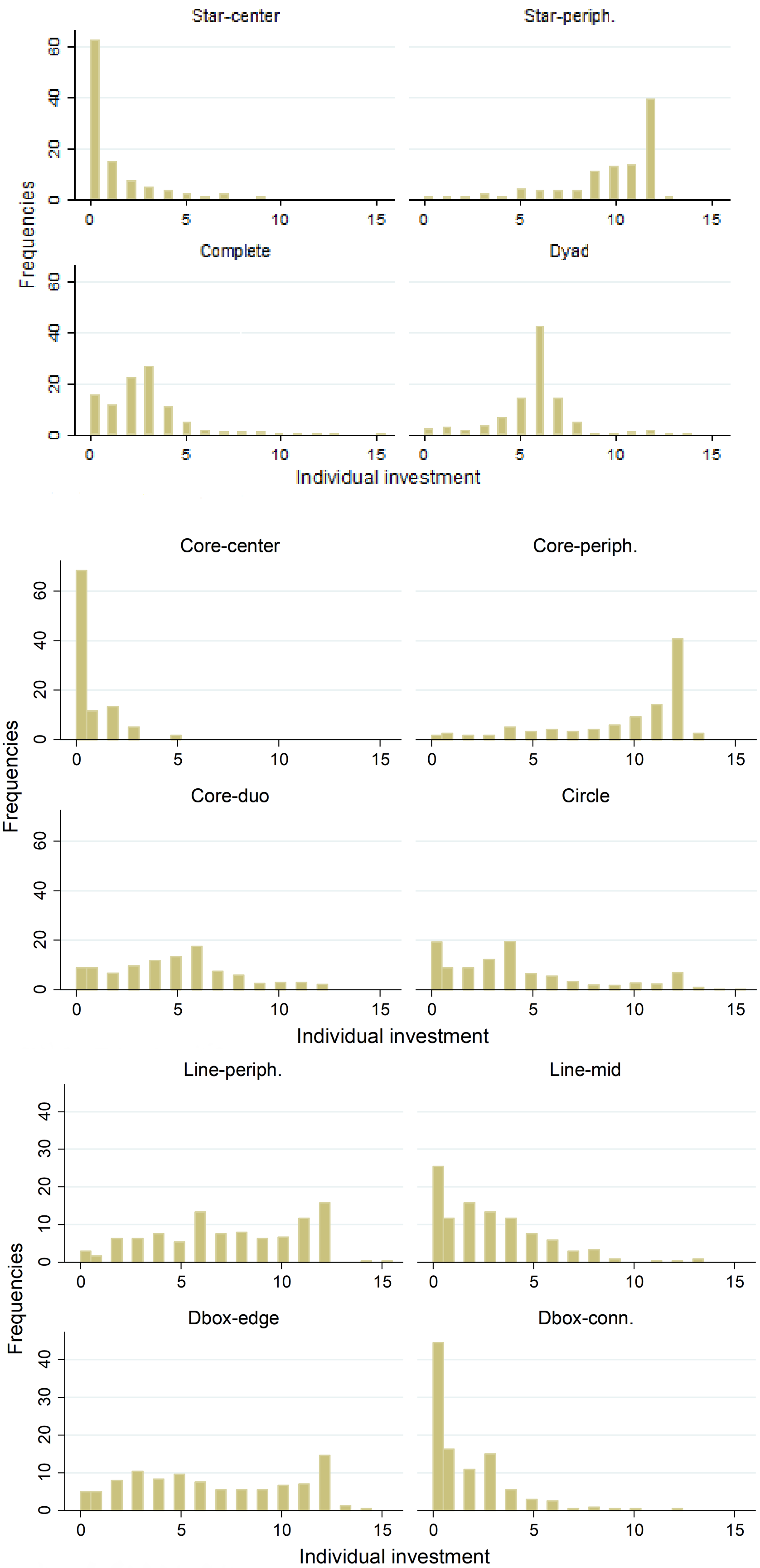
Nonetheless, this last statement requires further verification because it is clear that, in equilibrium, the investments of all players need to "fit" together. For this reason, we briefly describe the observed investment patterns at the network level, now. Table 1 presents the shares of investment profiles per network that are either consistent with the wider class of ORE, which we predicted for any group of socially concerned players, or the narrower class of refined ORE, which we just predicted for a group with compatible social preferences. The profiles are further classified based on their deviations from a standard Nash equilibrium for money-maximizing players, and we distinguish between zero ($\chi = 0$), two ($\chi < 3$), and any unit (any χ) of deviation from a standard best-response investment.[22]

We first have a look at column 3 (ORE with $\chi = 0$). There, we see that, on the asymmetric networks (star, core-periphery, d-box, and line), our earlier position-level findings are fully confirmed: Virtually all groups concluding their rounds with an ORE profile (52 in total) coordinated their investments on a periphery-specialized public good. The only exception to this rule is two groups playing the line network (1.6% of all groups playing the line), who coordinated on a distributed equilibrium where one of the end players earned more than her neighbor in the line middle.

Very similar patterns emerge for the dyad and the circle network. On the dyad, a large majority of groups (32.1%) converged on an equal-split equilibrium, a pattern that was already visible in Figure 4. Moreover, on the circle, we observe the same dispersed investment pattern that we already saw before: 7.5% of groups playing this network converged on a specialized equilibrium, that is, a profile with $(12, 0, 12, 0)$

[22] The exact criteria for the Nash equilibria with money-maximizing players and our (refined) ORE predictions are outlined in Table 7 in Appendix A. We chose a critical value of $\chi < 3$ because a deviation of up to two units is the maximum deviation for which a periphery-specialized public good emerges as the unique refined ORE in all asymmetric networks (except in the d-box, where the critical value is already at $\chi < 2$).

Figure 4: Investments by network position



NOTES: Observations from random round ends in the 960 payoff-relevant rounds of the experiment. One investment in the dyad with value 29 dropped for better display.

Table 1: Frequencies of other-regarding equilibria

Network	Equilibrium type	Deviation from money-maximizing equilibrium zero ($\chi = 0$)	moderate ($\chi < 3$)	any (any χ)
Dyad	equal split (rfd)	32.1%	45.8%	49.2%
	other	8.8%	33.0%	50.8%
Complete	equal split (rfd)	0.8%	0.8%	0.8%
	other	20.8%	62.5%	99.2%
Star	per-spec. (rfd)	15.8%	33.3%	62.5%
	distr. with $\pi_c \geq \pi_j$ (rfd)	—	—	36.6%
	cent-spec. or distr.	0%	0.8%	0.8%
Circle	specialized	7.5%	16.7%	29.2%
	distributed	3.3%	27.5%	55.0%
Core	per-spec. (rfd)	17.5%	43.3%	68.3%
	distr. with $\pi_c \geq \pi_j$ (rfd)	—	—	31.7%
	cent-spec. or distr.	0%	0%	0%
D-box	per-spec. (rfd)	8.3%	15.0%	25.8%
	distr. with $\pi_c \geq \pi_j$ (rfd)	—	1.7%	64.2%
	cent-spec. or distr.	0%	9.2%	10.0%
Line	end-spec. (rfd)	0.8%	40.1%	49.2%
	distr. with $\pi_m \geq \pi_e$ (rfd)	8.3%	13.3%	16.7%
	mid-spec. or distr.	1.6%	8.3%	34.1%

NOTES: Percentages of investment profiles consistent with an other-regarding equilibrium (ORE) at random round ends. 240 observations for the dyad, and 120 for all other networks. Refined OREs are indicated with "(rfd)".

or $(0, 12, 0, 12)$, and another 3.3% on a fully distributed equilibrium, with $(4, 4, 4, 4)$. Only on the complete network, merely one group (0.8%) reached the predicted equal-split equilibrium, as opposed to 25 groups (20.8%) who ended their rounds with an uneven distribution of a total investment of twelve units. Nonetheless, even this low share of refined ORE is not entirely surprising from the viewpoint of our theory. As we hypothesized, the size of the complete network renders coordination a challenging task, in particular for a group of players who differ in their social preferences (see Hypothesis 2).

Columns 4 and 5 tell a similar story. There, we look at the wider classes of ORE where some deviations from a money-maximizing equilibrium are considered consistent with our theory as well as long as these deviations are in line with the conditions in (3)–(6) . In the star, core, d-box, and line, the vast majority of investment profiles (87% across the four networks) were either consistent with a periphery-specialized ORE or with a distributed ORE profile where, however, the center player earned more than at least one peripheral player. Similarly, on the dyad, 49.2% of all round-end investment profiles were consistent with our predicted equal-split equilibrium, while on the circle, the shares of specialized and distributed investment profiles remain both on a high level. In sum, thus, the numbers in Table 1 largely confirm our theoretical predictions.[23]

5.2 Test of Hypothesis 1: The role of preference compatibility

So far, we have seen that many subject groups in our experiment coordinated their choices on a refined ORE. Nevertheless, this observation does not apply to all subject groups alike because there was also a sizable number of groups that failed to converge to a refined ORE, in particular when we look at OREs in the narrower sense (with $\chi < 3$). That heterogeneity is, however, part of our theory as well. Specifically, Hypothesis 1 posited that coordination is only easy to achieve for groups with compatible social preferences. Groups with incompatible preferences have, in contrast, a much harder time to coordinate because they face the same large number of equilibria to coordinate on as the players in the original game.

Social preference estimates: To test this hypothesis, we first needed to estimate the social preferences of our subjects. We did this based on the estimation strategy outlined in Section 4.2. Table 2 summarizes the resulting point estimates and categorizes them into their revealed preference types.[24] Consistent with the findings from earlier experiments (e.g., Bruhin et al., 2019; Kerschbamer and Müller, 2020), the table indi-

[23]To put these findings into perspective, we also compared the predictive power of our theory with that of several alternative equilibrium refinement concepts that were previously applied to the Bramoullé and Kranton game, notably efficiency, Nash tâtonnement stability, and quantal-response equilibrium. Our findings are detailed in Appendix B.1. To sum them up here, our key finding is that our refined ORE concept predicts the observed investment profiles at least as well as any of the alternative refinement concepts across all network structures investigated in our experiment. The specific power of our theory is that it selects the "natural" equilibria on most of these networks, such as an equal-split equilibrium on the dyad and a periphery-specialized equilibrium on the star, core-periphery, and d-box. At the same time, it can explain the observed behavioral heterogeneity on the line and the circle network, something that the other concepts are incapable of.

[24]The somewhat lengthy Table 9 that also categorizes our estimated $(\hat{\rho}_i, \hat{\sigma}_i)$-pairs into their revealed preference strengths is relegated to Appendix A.3.

Table 2: Social preference types

Preference type	Share
altruism ($\hat{\rho}_i \geq \hat{\sigma}_i > 0$)	11.7%
social welfare ($\hat{\rho}_i > \hat{\sigma}_i = 0$)	15.0%
inequity averse ($\hat{\rho}_i > 0 > \hat{\sigma}_i$)	29.2%
competitive ($0 = \hat{\rho}_i > \hat{\sigma}_i$)	10.0%
spiteful ($0 > \hat{\rho}_i \geq \hat{\sigma}_i$)	23.3%
payoff maximizer ($\hat{\rho}_i = \hat{\sigma}_i = 0$)	4.2%
asocial ($\hat{\sigma}_i > 0 > \hat{\rho}_i$)	6.7%
	100.0%

NOTES: Categorization of estimated $(\hat{\sigma}_i, \hat{\rho}_i)$-pairs into their revealed preference types. Insignificant estimates (i.e., p-values ≥ 0.05) or estimates with $-0.05 \leq x \leq 0.05$ for $x \in \{\hat{\sigma}_i, \hat{\rho}_i\}$ are set to zero because a subject with such a small preference parameter would make a decision indistinguishable from a payoff maximizer in our experiment.

cates substantive between-subject heterogeneity in social preferences. In particular, there was a sizable number of subjects in our experiment who displayed a behavior consistent with each one of the preference types identified by utility model (2).

Preference compatibility: In the next step, we used our estimates to classify all subject groups playing one of the six nested networks in our experiment into whether their members met the network-specific preference compatibility requirements or not. The criteria are outlined in Sections 3.4.1–3.4.3, and our classification results are presented in Table 3.[25] Clearly, the table indicates that, for at least four of the six networks, there was a substantial number of groups that had the required preference combination, while another large number did not have it. As expected as well, the share of groups with compatible preferences go down, sometimes quite substantively, when we focus on those groups that displayed small or moderate social preferences.[26] With this classification at hand, we can thus turn to our main question.

[25]To verify the additional requirement regarding the relative preference strengths of the players in the complete network, we demanded that all four subjects must exhibit either small ($\hat{\epsilon} < 1$), moderate ($1 \leq \hat{\epsilon} < 3$), or large ($3 \leq \hat{\epsilon}$) social comparison concerns.

[26]Interestingly also, the cross-network variation in the numbers of groups with compatible social preferences is much in line with the frequencies of refined ORE observed in Table 1. This already indicates that preference compatibility is a major driver behind the observed pattern of play.

Table 3: Preference compatibility and refined ORE

Networks Groups	Dyad	Star	Core	D-box	Line	Complete
Any preference strength (any $\hat{\epsilon}$):						
No. of groups	240	120	120	120	120	120
Groups with compatible preferences	31.3%	80.0%	40.0%	4.2%	26.7%	4.2%
Refined ORE share (any χ):						
Compatible groups	61.2%**	99.8%**	100%**	100%**	77.9%**	0%
Incompatible groups	41.4%	95.8%	98.9%	88.8%	70.2%	1.7%
Moderate preference strength ($\hat{\epsilon} < 3$):						
No. of groups	240	48	49	22	49	120
Groups with compatible preferences	31.3%	75.0%	42.9%	0%	28.6%	4.2%
Refined ORE share ($\chi < 3$):						
Compatible groups	57.9%**	27.1%	34.6%**	—	57.2%**	0%
Incompatible groups	35.9%	25.7%	28.8%	8.2%	42.6%	1.4%
Refined ORE share ($\chi = 0$):						
Compatible groups	40.1%**	15.7%**	11.8%	—	10.5%**	0%
Incompatible groups	22.7%	7.3%	10.3%	8.2%	2.2%	1.4%

NOTES: Preference compatibility and shares of refined ORE for all subject groups playing a nested network. Refined ORE shares are separately shown for groups with compatible and incompatible social preferences: ** indicates a significant difference at $p < 0.05$.

Hypothesis test: Do groups with compatible social preferences play a refined ORE more often than groups without the required preference combination? To answer this question, we refer to Table 3 again. There, we also contrasted the shares of refined ORE played by groups with and without a compatible preference combination respectively, as percentage shares of their total number of payoff-relevant investment decisions. The results generally support Hypothesis 1: While preference compatibility does not guarantee refined ORE play, it clearly facilitated coordination on these profiles. In the top panel of Table 3, which compares refined ORE profiles in the widest sense (any χ), the difference is still hardly visible for three of the six networks.[27] Nevertheless, when we focus on refined ORE in a narrower sense (with $\chi < 3$ or $\chi = 0$)——and consequently narrow our sample to groups with moderate social preference strengths ($\hat{\epsilon} < 3$)—, we

[27]This small gap is not surprising. As we already saw in Table 1, nearly all groups coordinated their choices on such a broadly defined ORE profile.

find a noticeable gap for four of the six nested networks.[28]

The systematic gap in refined ORE play can be further supported through multinomial logit regressions. Table 4 presents the coefficients and test statistics for two such models, both with the same dependent variable. The variable categorizes all conceivable investment profiles into six different outcome classes: Outcomes (1)–(3) capture our refined ORE predictions, while outcomes (4)–(6) encompass the remaining non-refined ORE.[29] Both outcome classes are further subdivided into the same deviations from a money-maximizing equilibrium that we already considered in Table 1.

The main explanatory variable in Model 1 is our social preference compatibility indicator. Model 2 further distinguishes between compatible groups with strong ($\hat{\epsilon} \geq 3$) and moderate ($\hat{\epsilon} < 3$) social preferences. Both models additionally include a number of control variables to account for several alternative factors that might explain why a certain investment profile is chosen more often than another. In particular, we included five network indicators to capture network-level factors and four group-level variables (gender, nationality, number of friends, and experience with the experimental game) to account for subject group characteristics that may be correlated with the social preferences of its members.

The results of these regressions lend further support to Hypothesis 1. The most compelling evidence comes from a series of post-estimation Wald tests following Model 1, where we test the impact of preference compatibility on various broader outcome classes. For instance, the Wald test (1–3) versus (rest) examines whether groups with compatible preferences played a refined ORE in the broadest sense (any χ) more often than any other profile. This is indeed confirmed, with a χ^2-statistic significant at the 0.01-level. The other two Wald tests concentrate on the refined ORE play in a narrower sense ($\chi < 3$). Consistent with our earlier observations, the results are more mixed here. As we already saw in Table 3, there was a considerable number of groups with strong social preferences in our experiment, and we would expect that these groups

[28] The two exceptions are the d-box and the complete network, where a meaningful comparison was simply impossible due to the limited number of groups meeting the demanding compatibility and preference-strength criteria for these networks.

[29] All other out-of-equilibrium profiles are subsumed under outcome (6).

also deviate significantly from a money-maximizing equilibrium.

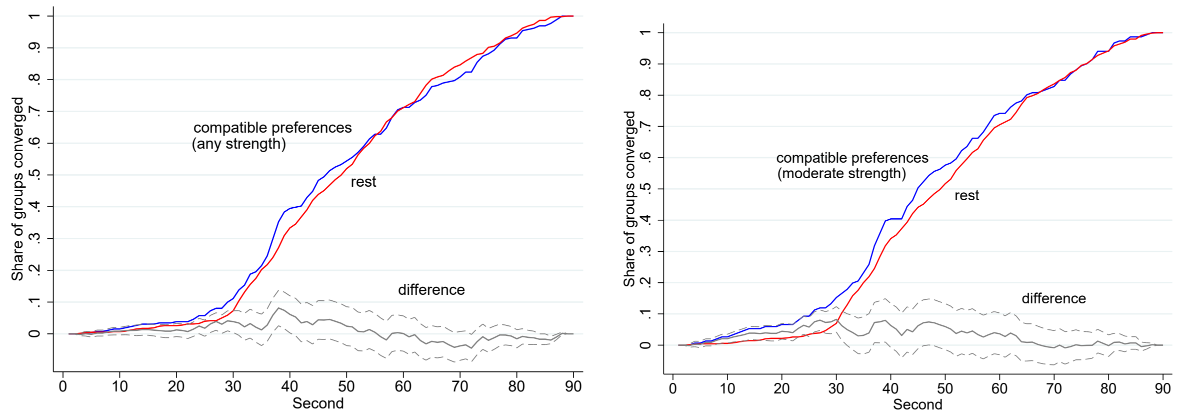
Table 4: Test of Hypothesis 1—Multinomial logit estimations

	Refined ORE			Non-refined ORE		
	$\chi = 0$	$0 < \chi < 3$	$3 \leq \chi$	$\chi = 0$	$0 < \chi < 3$	$3 \leq \chi$
	(1)	(2)	(3)	(4)	(5)	(6)
Model 1:						
Compatibility	0.92	0.78	0.81	-0.91	0.19	base
	(0.32)	(0.31)	(0.31)	(0.45)	(0.27)	outcome
Wald test of *Compatibility=0*:						
(1) versus (rest)	4.08**					
(1–2) versus (rest)		6.02**				
(1–3) versus (rest)			13.53***			
Model 2:						
Compatibility ($\hat{e} \geq 3$)	0.17	0.59	0.52	-11.73	-1.05	—
	(0.72)	(0.67)	(0.65)	(0.89)	(0.89)	
Compatibility ($\hat{e} < 3$)	1.07	0.76	0.78	-0.81	0.29	—
	(0.33)	(0.31)	(0.32)	(0.45)	(0.28)	
Wald tests:						
Compatibility ($\hat{e} \geq 3$)=0						
(1) versus (rest)	0.33					
(1–2) versus (rest)		0.38				
(2–3) versus (rest)			2.95*			
Compatibility ($\hat{e} < 3$)=0						
(1) versus (rest)	6.93***					
(1–2) versus (rest)		8.54***				
(2–3) versus (rest)			0.76			

NOTES: Coefficients and standard errors (clustered at group level) of two multinomial logit models. 24,299 observations from all payoff-relevant decision moments ($t \in [30, t^{max}]$) in all networks but the circle. Models include five unreported network indicators, seven session indicators, group characteristics (same sex, same nationality, number of friends), and two measures of group experience: general experience with our experiment (measured by the round number in a session) and experience with the current network (measured by the number of repetition). Wald tests report χ^2-statistics: $^{***}p < 0.01$,** $p < 0.05$,* $p < 0.1$.

To assess whether these groups also behaved as predicted, we developed Model 2. In line with our expectations, the associated post-estimation Wald tests validated that groups with compatible and strong social preferences (compatibility ($\hat{e} \geq 3$)), indeed, tended to coordinate on the broader class of ORE, while groups with compatible and weaker social preferences leaned towards the narrower ORE profiles. Thus, in sum, our findings from this section largely support our theoretical predictions that, within nested networks, preference compatibility enhances group coordination on a refined ORE. Combined with our earlier findings, they furthermore suggest that primarily the

Figure 5: Preference compatibility and time to convergence



NOTES: Shares of groups within the six nested networks converging already at time t to their final investment profile at t^{max}. Shares are shown separately for groups with compatible and incompatible preferences. Gray solid lines indicate between-group differences, and gray dashed lines their 90% confidence intervals.

groups with compatible social preferences played the most frequently observed refined ORE profiles in our experiment.

Time to convergence: Based on the above, one might wonder whether preference compatible also has an impact on the time to coordination. Even though not explicit part of our theory, it is intuitively plausible that a shared understanding of the expected investment profile reduces the time a group needs to coordinate its choices.

This question is examined in Figure 5, which investigates, for all the six nested networks combined, how many groups reached already at time t the final investment profile they will play in t^{max}. Clearly, the figure supports the expected accelerating impact of preference compatibility, in particular in the middle of a round between 30 and 50 seconds (left panel). The advantage becomes even more pronounced for groups with at most moderate social preferences ($\hat{e} < 3$), where the difference is already visible as early as 10 seconds after a round commenced (right panel). From the viewpoint of our theory, this is not surprising because we even predicted a unique ORE for these groups.

Discussion: Undeniably, the outcomes presented above ultimately depend on the precision of our social preference estimates. For pragmatic reasons, we chose an in-game measurement of these estimates. Consequently, they may be confounded by some other social preference concerns of our participants, such as their concern for reciprocity (Charness and Rabin, 2002; Dufwenberg and Patel, 2017). It is important to note, however, that any imprecision at this stage only works against us because it introduces measurement error into our preference compatibility indicator. In other words, we likely incorrectly classified several subject groups as having the right or wrong preference combination for a certain network. However, such misclassifications only introduce downward bias in our estimate for the true effect of preference compatibility because groups that truly had an easy time coordinating their choices might have been mistakenly mixed up with those facing greater coordination challenges, and vice versa.

Nevertheless, despite this potential source of error, we also have good reasons to believe that its impact on our findings is relatively mild. One part of the reason is that in all the asymmetric networks (star, core, d-box, line), a wide range of social preference types is compatible with the requirements in the critical central positions. This renders it unlikely that we have erroneously classified a substantive share of our participants as having incompatible preferences for these positions. Moreover, it is equally unlikely that we have misclassified a significant number of subject groups with genuinely incompatible preferences as being compatible with the dyad or the complete network. The stringent requirements for these networks would necessitate major measurement errors for multiple group members. Thus, we expect our above results to be quite robust.

To substantiate this claim, we conducted further sensitivity checks. For this purpose, we drew on social preference estimates from prior experimental studies on comparable student populations (Fehr and Charness, 2023) and simulated the impact of various degrees of measurement error in our own estimates. The results of these checks, detailed in Appendix B.3, indicate that any additional 10% chance of measurement error at the individual level reduces, on average across all networks, the effect of preference compatibility on refined ORE shares by 3 percentage points. For a sizable

Table 5: Placebo test on the circle

Preference strength	Strong ($\hat{\epsilon} \geq 3$)		Weak ($\hat{\epsilon} < 3$)	
Compatibility requirements	Complete	Line	Complete	Line
No. of groups	103	103	17	17
Groups with compatible preferences	2.9%	27.2%	11.7%	35.3%
Refined ORE share:	widest (any χ)		narrow ($\chi < 3$)	
Compatible groups	100%**	9.8%	53.1%	4.3%
Incompatible groups	52.6%	10.4%	60.7%	0%

NOTES: Preference compatibility and shares of refined ORE for all subject groups playing the circle network. Shares are shown separately for groups with preference combinations that are (are not) compatible with the criteria for the complete network or the line: ** indicates a significant difference at $p < 0.05$.

measurement error chance of 30%, for instance, this amounts to an effect reduction from 17 to still 8 percentage points, so a distorting effect well below the one found in other contexts (e.g. Gillen et al., 2019).

5.3 Test of Hypothesis 2: The impact of network nestedness

Our second hypothesis posited significant differences in the capacity of our experimental networks to promote coordination. More concretely, we conjectured that a refined ORE profile would be reached more easily within the nested networks, particularly in those networks where a single player nests the neighborhoods of all the other players. Two pieces of evidence support this conjecture.

Placebo test on the circle: According to our theory, social preferences should not facilitate group coordination on any of the three Nash equilibrium profiles on the circle network, even not when all four players are inequity-averse, competitive, or social-welfare types. The intuition here is that, due to the absence of nested neighborhoods in this network, players are unlikely to reach consensus on which equilibrium to play. Only in cases where all four players possess strong social preferences can an impact of their preference types be expected. Examples 3 and 4 illustrate this point most clearly.

Table 6: Frequency of refined ORE per network

Shares of refined ORE	Star	Core	D-box	Line	Complete
any χ:	0.99	0.99	0.89	0.72	0.02
$\chi < 3$:	0.29	0.30	0.17	0.52	0.01
$\chi = 0$:	0.12	0.10	0.08	0.08	0.01

NOTES: Data from all payoff-relevant decision moments in a network. All between-network differences of size $|d| > 0.01$ are statistically significant in two-sided t-tests at $p < 0.05$.

To put this to a test, we thus sought subject groups playing the circle network with a preference combination that proved effective in other networks. Specifically, we looked for subject groups whose preferences aligned with the compatibility requirements for the complete network and asked whether these were the groups that played the frequently observed distributed ORE profiles on the circle. Similarly, we searched for groups that matched the preference requirements for the line network and examined whether they were more likely to play the equally frequent specialized OREs on the circle. A systematic relationship should only be expected when groups exhibited strong social preferences in addition. Our findings, summarized in Table 5, support this assertion. The figure suggests no systematic relationship between preference types and group behavior when groups had at most moderate social preferences, but a noticeable relationship when groups exhibited strong social preferences. Most notably, groups that matched the preference compatibility requirements for the complete network almost always played the predicted distributed ORE profile.

Network comparisons: Our second piece of evidence comes from a comparison of the refined ORE shares across different networks. Hypothesis 2 predicted a negative impact of the complete network's size but a positive impact of the degree of nestedness, especially when all players' neighborhoods are nested within the neighborhood of a single player, such as in the star or core-periphery network. As we already saw in Section 5.1, our subjects evidently coordinated much more often on a refined ORE in the dyad than in the complete network, so this aspect is supported. Regarding the role of nestedness, Table 6 reproduces the shares of refined OREs for the star, core-periphery, d-box, line, and complete network. Consistent with our hypothesis, the shares are highest in the star and core-periphery, intermediate in the d-box and line, and lowest in the complete network. With a single exception, this ranking can also be confirmed in two-sided t-tests.[30]

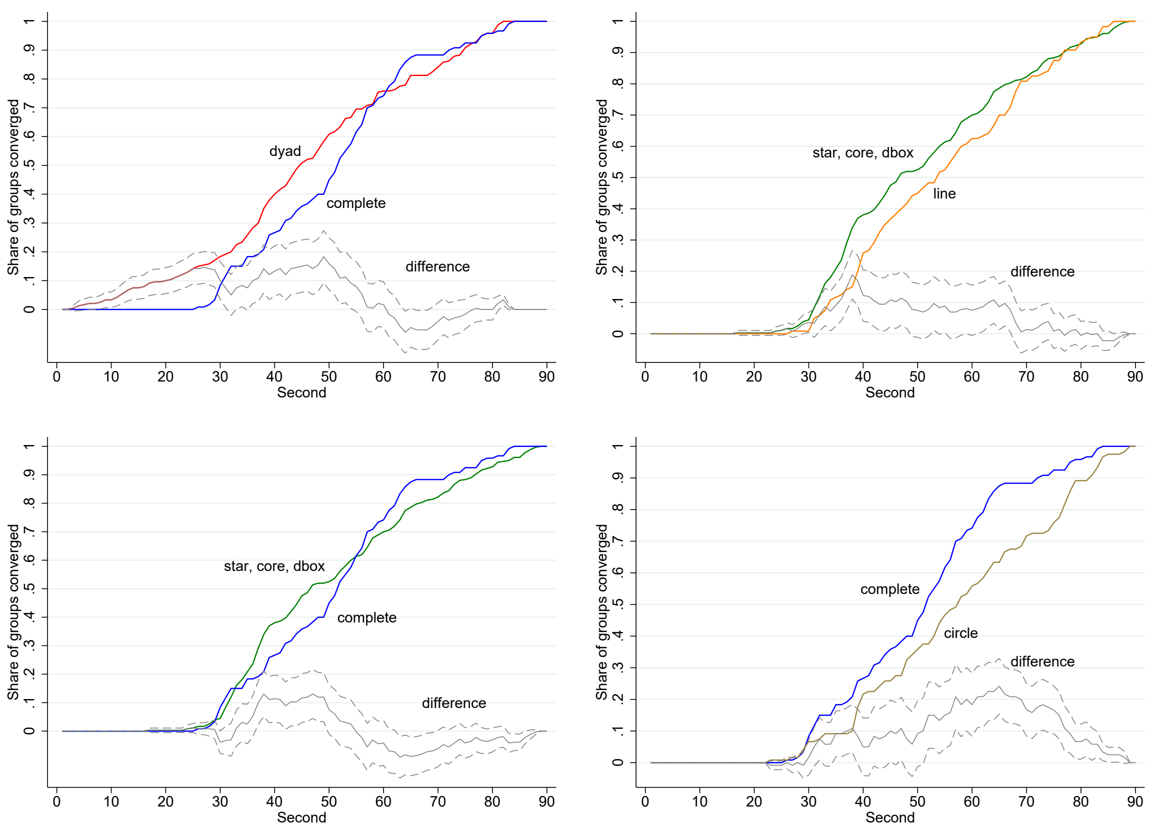
Time to convergence: Similar to the previous section, it also seems plausible to ask whether the structure of a network has an impact on the time a group needs to coordinate its choices. This is looked at in Figure 6, where the upper left panel clearly confirms the detrimental impact of the complete network's size. In contrast, the other three panels illustrate how a network's nestedness helps to expedite the time to coordination. The most demonstration of this can be found in the lower right panel, where it becomes evident that a group takes much longer to coordinate on the circle network compared to even the complete network. Altogether, thus, our findings in this section strongly support the second key prediction of our theory that coordination is more readily achieved when a network is nested.

6 Implications

Social preferences are a powerful trait in human behavior that can help foster cooperation, increase public goods provisions, or establish norms of good behavior. Social networks, by contrast, impose a constraint on the feasible distributions of the resulting gains and costs within a group or society. Our theory and experiment indicated that, in the realm of public goods provisions in small-scale networks, the network constraint prevails. In particular, we found that a group of socially concerned players does not deviate much from the behavior that also a group of pure maximizers would display. Nevertheless, when players' social preferences align with their network positions, they coordinate their behavior more easily and more swiftly on one of the game's equilibria.

[30]The exception is the ordering observed for the star and the line network, which deviates from our predictions when we consider the class of refined ORE in a narrower sense ($\chi < 3$), but nor when we consider refined ORE in the narrowest sense ($\chi = 0$).

Figure 6: Time to convergence per network



NOTES: Shares of groups per network converging already at time t to their final investment profile at t^{max}. Solid gray lines indicate between-network differences, gray dashed lines their 90% confidence intervals. Shares for star, core-periphery, and d-box are merged because of small within-differences.

In the following, we discuss two domains where these insights find practical applications: the organization of co-worker teams and public goods provisions in larger networks.

6.1 Organization of co-worker teams

Teams have gained increasing prominence in work settings, especially within knowledge-intensive organizations (e.g., Jones, 2021; Jarosch et al., 2021). Our findings provide insights int the optimal management of such knowledge-intensive organizations, where workers typically engage in multiple teamwork projects, and the network of teams shapes the spillovers between workers.

In these settings, managers often lack the means to enforce individual efforts from workers. However, they can influence the spillover network by, for instance, fixing the reporting lines or creating co-working spaces. Moreover, they can appoint the ideal candidates to the various positions in the network, taking (proxies for) their social preferences into account. Our findings highlight the importance of the suitable social preference types for each position within a network. Depending on the managers' goals, different network structures and types of workers are optimal. For instance, if the primary goal is to maximize the total effort of all workers in the shortest possible time, a star network with a competitive or spiteful type in the center position is the ideal form. On the other hand, if the aim is to maintain "social peace" and achieve the highest total effort under a fair distribution of inputs, then a complete network with inequity-averse or social-welfare types in each position should be the preferred choice.

6.2 Public goods in larger networks

Our empirical findings initially speak to the behavior in small-scale network games, such as the game implemented in our experiment. However, informed by our theory, they might empower us to speculate about the behavior in the larger social interaction networks that motivated our study. Can we expect social preferences help to eradicate the (in-)equality that is inherent in the structure of these network? In particular, do social preferences support more equitable payoff distributions when a network itself is asymmetric?

Proposition 2 allows us to make a clear-cut prediction under two conditions: first, the individuals reside in the same nested neighborhood of a network, and second, their social preferences align with their respective network positions. In this case, the individuals' payoffs and behaviors should simply reflect their centrality in the network. In other words, the social preferences of these individuals merely reinforce the inequality that is already inherent in the network structure.

In fact, the first of the two conditions is satisfied in many contexts. Nestedness is a well-documented topology of many social networks (Mariani et al., 2019) and it emerges as the outcome of various network formation processes (e.g., König et al., 2014). While little is known about the distribution of social preferences in networks, our theory suggests a stronger connection between the structure of a network and the distribution of payoffs in a more homogeneous (e.g., same-sex, same-age) group or society with shared social preferences.

7 Conclusion

We set out to study how social preferences shape behavior in a complex network game with multiple equilibria. Towards this end, we endowed the players in the local public goods game of Bramoullé and Kranton (2007) with social preferences and conducted an experiment to test our game's predictions. The results largely confirm the key prediction from our theory that social preferences can facilitate coordination on specific investment profiles, given the players' networks are mutually nested and their social preferences are compatible with their respective positions in the network.

As suggested by our theory, the key mechanism underlying our findings is that preference compatibility fosters a common understanding among players regarding which equilibrium to play. In the small-scale networks of our experiment, numerous player groups indeed appeared to share this common understanding. However, the question remains whether the same logic also extends to larger networks. We leave this question for future studies to explore.

References

Allouch, N. (2015). On the private provision of public goods on networks. *Journal of Economic Theory*, 157:527–552. 2

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review*, pages 891–904. 4.1

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636. 1, 3

Berninghaus, S. K., Ehrhart, K.-M., and Keser, C. (2002). Conventions and local interaction structures: experimental evidence. *Games and Economic Behavior*, 39(2):177–205. 3, 2

Berninghaus, S. K., Ehrhart, K.-M., and Ott, M. (2006). A network experiment in continuous time: The influence of link costs. *Experimental Economics*, 9(3):237–251. 4

Binmore, K. (2005). *Natural justice*. Oxford University Press. 2

Boncinelli, L. and Pin, P. (2012). Stochastic stability in best shot network games. *Games and Economic Behavior*, 75(2):538–554. 2

Bourlès, R., Bramoullé, Y., and Perez-Richet, E. (2017). Altruism in networks. *Econometrica*, 85(2):675–689. 2, 7

Bramoullé, Y. and Kranton, R. (2007). Public goods in networks. *Journal of Economic Theory*, 135(1):478–494. 1, 2, 3.1, 3.4.6, 4, 7, B.1

Bramoullé, Y., Kranton, R., and D'Amours, M. (2014). Strategic interaction and networks. *American Economic Review*, 104(3):898–930. 2

Bruhin, A., Fehr, E., and Schunk, D. (2019). The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association*, 17(4):1025–1069. 1, 6, 17, 5.2, 38

Callander, S. and Plott, C. R. (2005). Principles of network development and evolution: An experimental study. *Journal of Public Economics*, 89(8):1469–1495. 4

Cassar, A. (2007). Coordination and cooperation in local, random and small world networks: Experimental evidence. *Games and Economic Behavior*, 58(2):209–230. 3

Charness, G., Feri, F., Meléndez-Jiménez, M. A., and Sutter, M. (2014). Experimental games on networks: Underpinnings of behavior and equilibrium selection. *Econometrica*, 82(5):1615–1670. 2, 37

Charness, G., Feri, F., Meléndez-Jiménez, M. A., and Sutter, M. (2023). An experimental study on the effects of communication, credibility, and clustering in network games. *Review of Economics and Statistics*, 105(6):1530–1543. 4

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869. 1, 3.2, 5.2

Choi, S. and Lee, J. (2014). Communication, coordination, and networks. *Journal of the European Economic Association*, 12(1):223–247. 4

Dasgupta, P. and Maskin, E. (1986). The existence of equilibrium in discontinuous economic games, i: Theory. *The Review of economic studies*, 53(1):1–26. 3.4

Dufwenberg, M. and Patel, A. (2017). Reciprocity networks and the participation problem. *Games and Economic Behavior*, 101:260–272. 2, 5.2

Eckel, C. C. and Harwell, H. (2015). Four classic public goods experiments: A replication study. *Replication in Experimental Economics*, page 13. 1

Eyster, E., Madarász, K., and Michaillat, P. (2021). Pricing under fairness concerns. *Journal of the European Economic Association*, 19(3):1853–1898. 2

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692. 1, 6

Fehr, E. and Charness, G. (2023). Social preferences: fundamental characteristics and economic consequences. CESifo Working Paper. 3.2, 5.2, B.3, 38

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994. 4.1

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868. 3.2

Fehr, E. and Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7):458–468. 2

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178. 4.1, C.1

Galeotti, A., Goyal, S., Jackson, M. O., Vega-Redondo, F., and Yariv, L. (2010). Network games. *The Review of Economic Studies*, 77(1):218–244. 2

Ghiglino, C. and Goyal, S. (2010). Keeping up with the neighbors: social interaction in a market economy. *Journal of the European Economic Association*, 8(1):90–119. 2, 7

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863. 5.2

Goyal, S., Rosenkranz, S., Weitzel, U., and Buskens, V. (2017). Information acquisition and exchange in social networks. *The Economic Journal*, 127(606):2302–2331. 2

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125. 4.1, C.1

Immorlica, N., Kranton, R., Manea, M., and Stoddard, G. (2017). Social status in networks. *American Economic Journal: Microeconomics*, 9(1):1–30. 2, 7

Jarosch, G., Oberfield, E., and Rossi-Hansberg, E. (2021). Learning from coworkers. *Econometrica*, 89(2):647–676. 6.1

Jones, B. F. (2021). The rise of research teams: Benefits and costs in economics. *Journal of Economic Perspectives*, 35(2):191–216. 6.1

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review*, pages 728–741. 2

Kerschbamer, R. and Müller, D. (2020). Social preferences and political attitudes: An online experiment on a large heterogeneous sample. *Journal of Public Economics*, 182:104076. 1, 5.2

König, M. D., Tessone, C. J., and Zenou, Y. (2014). Nestedness in networks: A theoretical model and some applications. *Theoretical Economics*, 9(3):695–752. 6.2

Mariani, M. S., Ren, Z.-M., Bascompte, J., and Tessone, C. J. (2019). Nestedness in complex networks: observation, emergence, and implications. *Physics Reports*, 813:1–90. 1, 6.2

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38. B.1

Reuben, E. and Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior*, 77(1):122–137. 1, 2

Richefort, L. (2018). Warm-glow giving in networks with multiple public goods. *International Journal of Game Theory*, 47(4):1211–1238. 2

Rosenkranz, S. and Weitzel, U. (2012). Network structure and strategic investments: An experimental analysis. *Games and Economic Behavior*, 75(2):898–920. 2, 4, B.1

Schulz, U. and May, T. (1989). The recoding of social orientations with ranking and pair comparison procedures. *European Journal of Social Psychology*, 19(1):41–59. 3.2

Zhang, Y. and He, L. (2021). Theory and experiments on network games of public goods: inequality aversion and welfare preference. *Journal of Economic Behavior & Organization*, 190:326–347. 4

A Theory Appendix

A.1 Existence of other-regarding equilibrium (ORE)

Proof of Proposition 1. We verify that the modified Bramoullé and Kranton game, featuring socially concerned players, satisfies the sufficient conditions for the existence of a pure-strategy Nash equilibrium by Debreu, Glicksberg and, Fan: convexity and compactness of the strategy space, along with continuity and quasiconcavity of the utility function.

Obviously, $[0, \bar{e}]$ is convex and compact. Moreover, utility function (2) is continuous for all $e = (e_i)_{i \in N}$. It remains to show that $U_i(e)$ is also strictly quasiconcave in e_i. Since $U_i(e)$ is differentiable almost everywhere, this means that for all $e_{-i} \in [0, \bar{e}]^{n-1}$ and any two $0 \leq e_i' < e_i'' \leq \bar{e}$, we need that

$$U_i(e_i'', e_{-i}) \;\geq\; U_i(e_i', e_{-i}) \;\;\Rightarrow\;\; \frac{\partial U_i(e_i', e_{-i})}{\partial e_i} \;>\; 0\,, \tag{A.1}$$

whenever $U_i(\cdot)$ is differentiable at e_i'.[31]

To prove this, suppose, to the contrary, that $\frac{\partial U_i}{\partial e_i} \leq 0$ at some e_i'. We will show that then $U_i(e_i'', e_{-i}) < U_i(e_i', e_{-i})$. To show this, suppose, first, that no corner point (with $\pi_i = \pi_j$ for some $j \in R_i$) is passed when player i's investment is increased from e_i' to e_i''. Let $R_i^-(e_i)$ ($R_i^+(e_i)$) denote the sets of players who earn strictly more (less) than i at an investment $e_i \in [e_i', e_i'']$. Likewise, let $N_i^-(e_i)$ ($N_i^+(e_i)$) denote the subsets of i's neighbors who earn strictly more (less) than i at e_i. Under the assumption of a quadratic payoff function, we then get for any $e_i \in [e_i', e_i'']$:

$$\begin{aligned}
\frac{\partial^2 U_i}{\partial e_i^2} \;&=\; b''\left(1 - \rho_i \frac{|R_i^+(e_i)|}{|R_i|} - \sigma_i \frac{|R_i^-(e_i)|}{|R_i|}\right) + \frac{\rho_i}{|R_i|} \sum_{j \in N_i^+(e_i)} b'' + \frac{\sigma_i}{|R_i|} \sum_{j \in N_i^-(e_i)} b'' \\
&=\; b''\left(1 - \rho_i \frac{|R_i^+(e_i)| - |N_i^+(e_i)|}{|R_i|} - \sigma_i \frac{|R_i^-(e_i)| - |N_i^-(e_i)|}{|R_i|}\right) \\
&<\; 0\,,
\end{aligned} \tag{A.2}$$

because $b'' < 0$ and $1 > \rho_i \geq \sigma_i > -1$. Hence, we get

$$U_i(e_i'', e_{-i}) - U_i(e_i', e_{-i}) \;=\; \int_{e_i'}^{e_i''} \frac{\partial U_i}{\partial x} dx \;<\; \frac{\partial U_i(e_i', e_{-i})}{\partial e_i}(e_i'' - e_i') \;\leq\; 0\,.$$

A contradiction to (A.1).

Next, suppose that we do pass a corner point (with $\pi_i = \pi_j$ for some $j \in R_i$) when we increase i's investment from e_i' to e_i''. Let $\hat{e}_i$ denote the first corner point to pass. Because $U_i(e)$ is strictly concave in e_i whenever it is differentiable (see (A.2)), we have $\frac{\partial U_i}{\partial e_i} < 0$ for all $\tilde{e}_i \in [e_i', \hat{e}_i)$. We will now show that it must then also be $\frac{\partial U_i}{\partial e_i} < 0$ for all $e_i > \hat{e}_i$.

To do so, let $R_i^-(\tilde{e}_i)$ ($R_i^+(\tilde{e}_i)$) denote, as before, the sets of players who earn strictly more (less) than i at $\tilde{e}_i$. Likewise, let $N_i^-(\tilde{e}_i)$ ($N_i^+(\tilde{e}_i)$) denote the sets of i's neighbors who earn strictly more (less) than i at $\tilde{e}_i$. Moreover, let ΔR_i^+ (ΔR_i^-) denote the sets of players who migrate from $R_i^-(\tilde{e}_i)$ to $R_i^+(e_i)$ (respectively, from $R_i^+(\tilde{e}_i)$ to $R_i^-(e_i)$) at the corner point $\hat{e}_i$, and let ΔN_i^+ (ΔN_i^-) be similarly defined. Note first that at least one of the migration sets must, by definition of a

[31] Clearly, $U_i(e_i', e_{-i})$ is not differentiable whenever $\pi_i(e_i', e_{-i}) = \pi_j(e_i', e_{-i})$ for some $j \in R_i$. Nevertheless, in these cases, condition (A.1) must hold for all e_i in a small open neighborhood around e_i'.

corner point, be non-empty. Note next that at $\tilde{e}_i \in [e_i', \hat{e}_i)$ it must be

$$\frac{\partial \pi_i(\tilde{e}_i, e_{-i})}{\partial e_i} > (<) \frac{\partial \pi_j(\tilde{e}_i, e_{-i})}{\partial e_i}$$

for every j who migrates from $R_i^-(\tilde{e}_i)$ to $R_i^+(e_i)$ (respectively, from $R_i^+(\tilde{e}_i)$ to $R_i^-(e_i)$). Otherwise, j would not migrate. Note finally that it is possible to write $\frac{\partial \pi_j(e_i)}{\partial e_i} = \frac{\partial \pi_j(\tilde{e}_i)}{\partial e_i} + b''(e_i - \tilde{e}_i)$ for any $j \in N_i \cup \{i\}$ and $\frac{\partial \pi_j}{\partial e_i} = 0$ for any $j \in R_i \backslash N_i$. Altogether, this means that for any e_i larger than the first corner point $\hat{e}_i$ (and smaller than the second corner point) that

$$\begin{aligned}
\frac{\partial U_i(e_i)}{\partial e_i} &= \frac{\partial \pi_i(\tilde{e}_i)}{\partial e_i}\left(1 - \rho_i \frac{|R_i^+(\tilde{e}_i)| + |\Delta R_i^+| - |\Delta R_i^-|}{|R_i|} - \sigma_i \frac{|R_i^-(\tilde{e}_i)| - |\Delta R_i^+| + |\Delta R_i^-|}{|R_i|}\right) \\
&+ \frac{\rho_i}{|R_i|} \sum_{j \in N_i^+(\tilde{e}_i)} \frac{\partial \pi_j(\tilde{e}_i)}{\partial e_i} + \frac{\sigma_i}{|R_i|} \sum_{j \in N_i^-(\tilde{e}_i)} \frac{\partial \pi_j(\tilde{e}_i)}{\partial e_i} \\
&+ \frac{\rho_i - \sigma_i}{|R_i|}\left(\sum_{j \in \Delta R_i^+} \frac{\partial \pi_j(\tilde{e}_i)}{\partial e_i} - \sum_{j \in \Delta R_i^-} \frac{\partial \pi_j(\tilde{e}_i)}{\partial e_i}\right) \\
&+ \frac{\partial^2 U_i(e_i)}{\partial e_i^2}(e_i - \tilde{e}_i) \\
&= \frac{\partial U_i(\tilde{e}_i)}{\partial e_i} + \frac{\partial^2 U_i(e_i)}{\partial e_i^2}(e_i - \tilde{e}_i) \\
&- \frac{\rho_i - \sigma_i}{|R_i|}\left(\frac{\partial \pi_i(\tilde{e}_i)}{\partial e_i}|\Delta R_i^+| - \sum_{j \in \Delta R_i^+} \frac{\partial \pi_j(\tilde{e}_i)}{\partial e_i} - \frac{\partial \pi_i(\tilde{e}_i)}{\partial e_i}|\Delta R_i^-| + \sum_{j \in \Delta R_i^-} \frac{\partial \pi_j(\tilde{e}_i)}{\partial e_i}\right),
\end{aligned}$$

where $\partial^2 U_i(e_i)/\partial e_i^2$ denotes the expression in (A.2) evaluated at e_i. Because all three summands in the final two lines are negative (and at least one of them is strictly negative), we get $\frac{\partial U_i(e_i)}{\partial e_i} < 0$.

Applying the same argument to all further corner points to pass, we thus get more generally $\frac{\partial U_i(e_i)}{\partial e_i} < 0$ for any $e_i \in [e_i', e_i'']$ whenever $U_i(e)$ is differentiable. This, in turn, means that $U_i(e_i'', e_{-i}) - U_i(e_i', e_{-i}) = \int_{e_i'}^{e_i''} \frac{\partial U_i}{\partial x} dx < 0$ or, in other words, U_i is strictly quasiconcave in e_i. Moreover, it follows from here that player i possesses a unique best response on every e_{-i}. ■

A.2 ORE characterization

Here, we provide a complete characterization of set of other-regarding equilibria (ORE) for the seven networks in our experiment, as well as a partial characterization of this set for a general network structure. The detailed predictions for our experimental networks are also summarized in Table 7.

Table 7: Predictions

	Payoff-max. equilibria	ORE	Refined ORE
Dyad and complete	$\sum_{i\in N} e_i = 12$ (S,E) (Q: $e_i = e_j = \frac{12}{n}$)	$\sum_{i\in N} e_i \in [12 \pm \epsilon]$	$e_i = e_j \in [\frac{12\pm\epsilon}{n}]$ if $\tau_i \in T_c \cap T_p \ \forall i \in N$ and $\rho_i \approx \rho_j, \sigma_i \approx \sigma_j$ in complete network
Star	(i) $e_c = 0, e_p = 12$ (ii) $e_c = 12, e_p = 0$ (E: (ii) selected)	(i) $e_c = 0, e_p \in [12 \pm \epsilon]$ (ii) $e_c \in [12 - \frac{7\epsilon}{3}, 12 + \epsilon]$,	$\pi_c \geq \min_{p \in P}\{\pi_p\}$ if $\tau_c \in T_c$ and $\exists\, p \in P : \tau_p \in T_p$ If also $\epsilon < 3$: $e_c = 0, e_p \in [12 \pm \epsilon]$
Core periphery	(i) $e_c = 0, e_p = 12$, $\sum_{d \in D} e_d = 12$ (ii) $e_c = 12, e_{-c} = 0$ (S: (i) selected) (Q: (i) with $e_d = 6$) (E: (ii) selected)	(i) $e_c = 0, e_p \in [12 \pm \epsilon]$, $\sum_{d \in D} e_d \in [12 \pm \epsilon]$ (ii) $e_c \in [12 - \frac{7\epsilon}{3}, 12 + \epsilon]$, $\sum_{j \neq c} e_j \leq 4\epsilon$	$\pi_c \geq \min_{j \neq c}\{\pi_j\}$ if $\tau_c \in T_c$ and $\exists\, j \neq c$: $\tau_j \in T_p \backslash \{ineq.av., comp.\}$ If also $\epsilon < 3$: $e_c = 0, e_p \in [12 \pm \epsilon]$, $\sum_{d \in D} e_d \in [12 \pm \epsilon]$
D-box	(i) $e_c = 0, e_p = 12$ (E) (ii) $e_p = 0$, $\sum e_c = 12$ (E) (S,Q: (i) selected)	(i) $e_c = 0, e_p \in [12 \pm \epsilon]$ (ii) $\sum e_c \in [12 - 3\epsilon, 12 + \epsilon]$, $\sum e_p \leq 4\epsilon$	$\pi_c \geq \min_{p \in P}\{\pi_p\}$ if $\tau_c \in T_c \backslash \{welfare\} \ \forall c \in C$ and $\exists p \in P$: $\tau_p \in T_p \backslash \{ineq.av., comp.\}$ If also $\epsilon < 2$: $e_c = 0, e_p \in [12 \pm \epsilon]$
Line	(i) $e_{p_i} = 12, e_{c_i} = 0$, $e_{c_j} + e_{p_j} = 12$ (S) (ii) $e_{p_i} = 12, e_{c_i} = 0$, (Q) $e_{c_j} = 0, e_{p_j} = 12$ (iii) $e_{p_i} = 12, e_{c_i} = 0$, $e_{c_j} = 12, e_{p_j} = 0$ (E)	$\forall i : e_i + \sum_{j \in N_i} e_j \geq e^* - \epsilon$ If also $\epsilon < 4$: (i) $e_{p_i} \in [12 \pm \epsilon], e_{c_i} = 0$, $e_{c_j} + e_{p_j} \in [12 \pm \epsilon]$ (ii) $e_p \in [12 - 3\epsilon, 12 + \epsilon]$, $e_c \leq 2\epsilon$	$\pi_c \geq \pi_p$ and $e_c \leq e_p$ if $\epsilon < 3$ and $\tau_c \in T_c \ \forall c \in C$ and $\tau_p \in T_p \ \forall p \in P$
Circle	(i) $e_i = 0, e_{i+1} = 12$ (ii) $e_i = 4$ (S,E: (i) selected) (Q: (ii) selected)	$\forall i : e_i + \sum_{j \in N_i} e_j \geq e^* - \epsilon$ If also $\epsilon < 3$: (i) $e_i = 0, e_{i+1} \in [12 \pm \epsilon]$ (ii) $e_i \in [4 \pm \epsilon]$	

NOTES: (Other-regarding) equilibria for the seven networks in our experiment with payoff function (11) and $e^* = 12$. For comparison, the equilibria selected by several alternative refinement concepts are highlighted as well: (S) asymptotic stability, (Q) quantal response equilibria with marginal decision errors, (E) efficient equilibria.

A.2.1 Star, core-periphery, d-box

ORE set: Building on Definition 3 (preference strength), we show that an ORE on the star, core-periphery, or d-box must either result in a center-specialized, periphery-specialized, or distributed public good.

Suppose, first, that $e_c = 0$ for all players in the center position(s) $c \in C$ (*periphery-specialization*). A payoff maximizer in a periphery position $p \in P$ would then respond with $f_p(e_{-p}) = e^*$. By Definition 3, a socially concerned player responds with $e_p \equiv f_p(\tau_p, e_{-p}) \in [e^* \pm \epsilon_p]$, and two social concerned players in the duo positions of the core-periphery with $e_d \equiv f_d(\tau_d, e_{-d})$, where $\sum_{d\in D} e_d \in [e^* \pm \epsilon_d]$. Using $\epsilon \equiv \max\{\epsilon_p, \epsilon_d\}$, we immediately arrive at the investment boundaries in a periphery-specialized ORE.

Next, suppose that $e_c > 0$ for at least one $c \in C$ (*center-specialized* or *distributed*). By Definition 3, the best-response investments of socially concerned players in the center, periphery, and duo positions must satisfy

$$e_c \in [e^* - \sum_{j\neq c} e_j \pm \epsilon_c], \tag{A.3}$$

$$e_p \in [e^* - \sum_{c\in C} e_c \pm \epsilon_p], \tag{A.4}$$

$$\sum_{d\in D} e_d \in [e^* - e_c \pm \epsilon_d]. \tag{A.5}$$

It follows from (A.3) that $\sum_{i\in N} e_i \leq e^* + \epsilon_c$ and from (A.4) and (A.5) that $e_p + \sum_{c\in C} e_c \geq e^* - \epsilon_p$ and $e_c + \sum_{d\in \mathcal{D}} e_d \geq e^* - \epsilon_d$. In combination, this means that the periphery players in the star and d-box (except for one peripheral player p_1) jointly contribute at most

$$\begin{aligned} \sum_{j\in P\setminus\{p_1\}} e_j &= \sum_{j\in P} e_j + \sum_{c\in C} e_c - \big(\sum_{c\in C} e_c + e_{p_1}\big) \leq e^* + \max_{c\in C}\{\epsilon_c\} - \big(e^* - \max_{p\in P}\{\epsilon_p\}\big) \\ &= \max_{c\in C}\{\epsilon_c\} + \max_{p\in P}\{\epsilon_p\}. \end{aligned}$$

Drawing the same conclusion for any other periphery player p_2, we again get $\sum_{j\in P\setminus\{p_2\}} e_j \leq \max_{c\in C}\{\epsilon_c\} + \max_{p\in P}\{\epsilon_p\}$ and, thus, the total contribution received by the center player(s) is at most

$$\sum_{p\in P} e_p \leq \sum_{j\in P\setminus\{p_1\}} e_j + \sum_{j\in P\setminus\{p_2\}} e_j \leq 2(\max_{c\in C}\{\epsilon_c\} + \max_{p\in P}\{\epsilon_p\}). \tag{A.6}$$

Similarly, in the core-periphery, the periphery and duo players contribute at most

$$\begin{aligned} e_p &= \sum_{l\in N\setminus\{c\}} e_l + e_c - \big(e_c + e_p\big) \leq \epsilon_p + \epsilon_c, \\ \sum_{d\in D} e_d &= \sum_{l\in N\setminus\{c\}} e_l + e_c - \big(e_c + \sum_{d\in D} e_d\big) \leq \max_{d\in D}\{\epsilon_d\} + \epsilon_c. \end{aligned}$$

The total contribution received by the center player is thus at most

$$\sum_{d\in D} e_d + e_p < 2\epsilon_c + \epsilon_p + \max_{d\in D}\{\epsilon_d\}. \tag{A.7}$$

For the peripheral player(s), (A.4) implies that their total investment received is constrained from below by $\min_{p\in P}\{e_p\} + \sum_{c\in C} e_c \geq e^* - \max_p\{\epsilon_p\}$. Similarly, in the core-periphery, (A.5) implies that $\sum_{d\in D} e_d$ is constrained from below by $\min_{d\in D}\{\sum_{d\in D} e_d\} + e_c \geq e^* - \max_{d\in D}\{\epsilon_d\}$. Thus, the center players' investments in the star and d-box are larger than

$$\begin{aligned} \sum_{c\in C} e_c &\geq e^* - \max_{p\in P}\{\epsilon_p\} - \max\{\min_{p\in P}\{e_p\}\} \\ &\geq e^* - \max_{p\in P}\{\epsilon_p\} - \frac{2(\max_{c\in C}\{\epsilon_c\} + \max_{p\in P}\{\epsilon_p\})}{n - |C|}, \end{aligned} \tag{A.8}$$

where the lower bound in the second line is determined by a situation where all peripheral players equally share $2(\max_{c\in C}\{\epsilon_c\} + \max_{p\in P}\{\epsilon_p\})$. Similarly, in the core-periphery, define

$\epsilon_j \equiv \max\{\epsilon_p, \epsilon_d\}$. Then, the center's investment is larger than

$$e_c \geq e^* - \epsilon_j - \max\{\min\{e_p, \sum_{d\in D} e_d\}\} \geq e^* - \epsilon_j - \frac{2(\epsilon_c + \epsilon_j)}{n-1}\,. \tag{A.9}$$

Finally, (A.3) implies that the center player(s)' investment is smaller than

$$\sum_{c\in C} e_c \quad \leq \quad e^* + \max_{c\in C}\{\epsilon_c\}\,. \tag{A.10}$$

Together, conditions (A.6)–(A.10) define the investment boundaries in a center-specialized or distributed ORE.

Refined ORE on star: We show that when $\tau_c \in T_c$ for the center player c and $\tau_p \in T_p$ for at least one peripheral player p, then

$$\pi_c(e) \geq \min_{j\in N\setminus C}\{\pi_j(e)\}\,. \tag{A.11}$$

To see this, suppose that, contrary to (A.11), $\pi_c(e) < \pi_p(e)$ for all $p \in P$. For this to occur in an ORE, we require for the center player c and any periphery player p that their first-order conditions are satisfied:[32]

$$
\begin{aligned}
(i)\quad \frac{\partial U_c(e)}{\partial e_c} \;&=\; (b'_c - c)(1-\sigma_c) + \frac{\sigma_c}{3}\sum_{p\in P} b'_p \;=\; 0\\
(ii)\quad \frac{\partial U_p(e)}{\partial e_p} \;&=\; (b'_p - c)\big(1 - \rho_p\frac{|R_p^+|}{|R_p|} - \sigma_p\frac{|R_p^-|}{|R_p|}\big) + \frac{\rho_p}{|R_p|}b'_c \;\leq\; 0\,.
\end{aligned}
$$

Here, b'_c and b'_p are our shorthand notations for $b'(e_c + \sum_{p\in P} e_p)$ and $b'(e_p + e_c)$ respectively. Moreover, player p may either just compare with c (i.e., $|R_p| = |R_p^+| = 1$) or with some other peripheral players in addition (i.e., $|R_p| > 1$).

Now, it follows from $\tau_c \in T_c$ and $\tau_p \in T_p$ that $\sigma_c \leq 0$ and $\rho_p \geq 0$ with at least one inequality being strict. As a result, condition (i) implies $b'_c - c \geq 0$. And because $b'_p \geq b'_c > 0$, we also get $b'_p - c \geq 0$. Yet, this means that $\partial U_p/\partial e_p > 0$. A contradiction to the other necessary equilibrium condition (ii). We must therefore have $\pi_c(e) \geq \min_{j\in N\setminus C}\{\pi_j(e)\}$.

Refined ORE on core-periphery: We show that when $\tau_c \in T_c$ for the center player c and $\tau_j \in T_p\setminus\{\textit{inequity averse}, \textit{competitive}\}$ for at least one non-center player $j \neq c$, then payoff ranking (A.11) must apply in an ORE.

Suppose, to the contrary, that $\pi_c(e) < \pi_j(e)$ for all $j \neq c$. For this to arise in an ORE, we need to have for the center c and the player j with $\tau_j \in T_p$ that their first-order conditions are satisfied. By the same argument as for the star network, this cannot be true when j is a periphery player. When j is a duo player, the most ideal constellation for an ORE is the one where $\pi_j(e) = \pi_k(e)$ for $j, k \in D$. This means that the following conditions need to hold for

[32]Here, we have assumed that $\pi_p \neq \pi_l$ for all $l \in R_p$. Nevertheless, because $U_p(e)$ is continuous, a very similar first-order condition to (ii) must hold for some small $h > 0$ and all $e'_p \in (e_p, e_p + h)$.

some small $h > 0$ and any $e'_j \in (e_j,\, e_j + h)$:

$$(i) \quad \frac{\partial U_c(e)}{\partial e_c} \;=\; (b'_c - c)(1 - \sigma_c) + \frac{\sigma_c}{3} \sum_{i \neq c} b'_i \;=\; 0$$

$$(ii) \quad \frac{\partial U_j(e'_j, e_{-j})}{\partial e_j} \;=\; (b'_j - c)\big(1 - \rho_j \frac{|R_j^+|}{|R_j|} - \sigma_j \frac{|R_j^-|}{|R_j|}\big) + \frac{\rho_j}{2} b'_c + \frac{\sigma_j}{2} b'_k \;\leq\; 0\,.$$

However, when $\tau_c \in T_c$ and $\tau_j \in T_p \backslash \{inequity\ averse\,,\, competitive\}$, it follows that $\sigma_c \leq 0$ and $\rho_j \geq \sigma_j \geq 0$. Thus, we can again apply the same argument as for the star network to conclude that (i) implies $\partial U_j(e'_j, e_{-j})/\partial e_j > 0$. A contradiction to the necessary equilibrium condition (ii). In an ORE, $\pi_c(e) \geq \min_{j \in N \backslash C}\{\pi_j(e)\}$ therefore needs to hold.

Refined ORE on d-box: When $\tau_c \in T_c \backslash \{inequity\ averse,\, social\ welfare\}$ for both centers $c \in C$ and $\tau_p \in T_p \backslash \{inequity\ averse,\, competitive\}$ for at least one $p \in P$, then payoff ranking (A.11) must apply to both centers in an ORE.

Suppose, to the contrary, that for at least one $c_1 \in C$ it holds $\pi_{c_1}(e) < \pi_p(e)$ for both $p \in P$. For this to occur in an ORE, we require for the center c_1 and some periphery player p_1 that their first-order conditions are satisfied. In particular, one of the favorable equilibrium constellations is the one where the other center player c_2 earns more than p_1, i.e., $\pi_{c_1}(e) < \min\{\pi_{p_1}(e), \pi_{p_2}(e)\} < \pi_{c2}(e)$. This means that the following conditions need to apply:

$$(i) \quad \frac{\partial U_{c_1}(e)}{\partial e_{c_1}} \;=\; (b'_{c_1} - c)(1 - \sigma_{c_1}) + \frac{\sigma_{c_1}}{3} \sum_{i \neq c_1} b'_i \;=\; 0$$

$$(ii) \quad \frac{\partial U_{p_1}(e)}{\partial e_{p_1}} \;=\; (b'_{p_1} - c)\big(1 - \rho_{p_1} \frac{|R_{p_1}^+|}{|R_{p_1}|} - \sigma_{p_1} \frac{|R_{p_1}^-|}{|R_{p_1}|}\big) + \frac{\rho_{p_1}}{2} b'_{c_1} + \frac{\sigma_{p_1}}{2} b'_{c_2} \;\leq\; 0\,.$$

However, when $\tau_{c_1} \in T_c$ and $\tau_{p_1} \in T_p \backslash \{inequity\ averse,\, competitive\}$, we get $\sigma_{c_1} \leq 0$ and $\sigma_{p_1} \geq 0$ and, thus, by the same arguments as made for the star network, condition (i) implies $\partial U_{p_1}(e)/\partial e_{p_1} > 0$. A contradiction to an ORE.

The other favorable equilibrium constellation is the one where $\pi_{c1}(e) = \pi_{c2}(e) < \min\{\pi_{p_1}(e), \pi_{p_2}(e)\}$. For this to establish an ORE, we require for some small $h > 0$ and any $e'_{c_1} \in (e_{c_1} - h,\, e_{c_1})$:

$$(i) \quad \frac{\partial U_{c_1}(e'_{c_1}, e_{-c_1})}{\partial e_{c_1}} \;=\; (b'_{c_1} - c)\big(1 - \frac{2\sigma_{c_1}}{3} - \frac{\rho_{c_1}}{3}\big) + \frac{\sigma_{c_1}}{3} \sum_{p \in P} b'_p + \frac{\rho_{c_1}}{3} b'_{c_2} \;\geq\; 0$$

$$(ii) \quad \frac{\partial U_{p_1}(e)}{\partial e_{p_1}} \;=\; (b'_{p_1} - c)\big(1 - \rho_{p_1} \frac{|R_{p_1}^+|}{|R_{p_1}|} - \sigma_{p_1} \frac{|R_{p_1}^-|}{|R_{p_1}|}\big) + \frac{\rho_{p_1}}{2} \big(b'_{c_1} + b'_{c_2}\big) \;\leq\; 0\,.$$

However, when $\tau_{c_1} \in T_c \backslash \{inequity\ averse,\, social\ welfare\}$ and $\tau_{p_1} \in T_p$, we have $\sigma_{c_1} \leq \rho_{c_1} \leq 0$ and $\rho_{p_1} \geq 0$. Hence again, we again arrive at a contradiction between the two necessary ORE conditions. The payoff condition (A.11) thus needs to hold to both center players in the d-box.

Refined ORE with limited preference strength: The payoff ranking condition (A.11) even translates into an investment ranking when the social preference of all players are sufficiently weak. To see how, note that in a *center-specialized* or *distributed* equilibrium, the center's investment converges, by (A.8) and (A.10), to

$$\lim_{(\epsilon_p, \epsilon_c, \epsilon_d) \to (0,0,0)} \sum_{c \in C} e_c = e^*\,.$$

Moreover, the investments of the non-center players $j \in N \backslash C$ converge, by (A.6) and (A.7), to

$$\lim_{(\epsilon_p, \epsilon_c, \epsilon_d) \to (0,0,0)} e_j = 0\,.$$

Thus, there exist $\overline{\epsilon}^{star} = \overline{\epsilon}^{core} \equiv \max\{\epsilon_p, \epsilon_c, \epsilon_d\}$ and $\overline{\epsilon}^{dbox} \equiv \max\{\epsilon_p, \epsilon_c\}$ such that for any smaller ϵ condition $\pi_c(e) \geq \min_{j \in N \backslash C}\{\pi_j(e)\}$ cannot be fulfilled.

The critical values can be determined as follows: in a *center-specialized* or *distributed* equilibrium on the star or core-periphery, the center's payoff is, by (A.6) and (A.7), lower than

$$\pi_c(e) \leq b\big(e^*\big) - c\big(e^* - 4\overline{\epsilon}^{star}\big) \equiv \max \pi_c(e)\,.$$

Moreover, because the center invests, by (A.8) and (A.9), more than $e^* - (7\overline{\epsilon}^{star})/3$ and each non-center player less than $2\overline{\epsilon}^{star}$, the non-center players' payoffs are larger than

$$\pi_j(e) \geq b\big(e^* - \frac{7\overline{\epsilon}^{star}}{3}\big) \equiv \min \pi_j(e)\,.$$

Hence, the critical value is defined by the largest $\overline{\epsilon}^{star}$ to satisfy $\max \pi_c(e) < \min \pi_j(e)$ or equivalently,

$$c > \frac{b\big(e^*\big) - b\big(e^* - \frac{7}{3}\overline{\epsilon}^{star}\big)}{e^* - 4\overline{\epsilon}^{star}}\,.$$

On the d-box, the critical value is given as follows: in a *center-specialized* or *distributed* equilibrium, the minimum of the center players' payoffs is, by (A.6), smaller than

$$\min_{i \in C}\{\pi_i(e)\} \leq b\big(e^*\big) - c\frac{e^* - 4\overline{\epsilon}^{dbox}}{2} \equiv \max\{\min \pi_c(e)\}\,.$$

Furthermore, because the centers invest, by (A.8), jointly more than $e^* - 3\overline{\epsilon}^{dbox}$ and each periphery player less than $2\overline{\epsilon}^{dbox}$, the peripherals' payoffs are larger than

$$\pi_p(e) \geq b\big(e^* - 3\overline{\epsilon}^{dbox}\big) \equiv \min \pi_p(e)\,.$$

Hence, the critical value is defined by the largest $\overline{\epsilon}^{dbox}$ to satisfy $\max\{\min \pi_c(e)\} < \min \pi_j(e)$ or equivalently,

$$c > \frac{b(e^*) - b\big(e^* - 3\overline{\epsilon}^{dbox}\big)}{e^* - 4\overline{\epsilon}^{dbox}}\,.$$

A.2.2 Line

ORE set: We show that an ORE must either entail an end-specialized or a distributed investment profile, given that players' social preferences are sufficiently weak.

Fix the sequence of players in the order $p1, c1, c2$, and $p2$, and suppose that all players possess small social preferences, in particular $\epsilon \equiv \max\{\epsilon_c, \epsilon_p\} < e^*/3$. Then, all ORE fall into one of the following two classes:

$$\begin{aligned} (\textit{end-sponsored}): &\quad \big([e^* - 3\epsilon, e^* + \epsilon]\,,\, [0, 2\epsilon]\,,\, [0, 2\epsilon]\,,\, [e^* - 3\epsilon, e^* + \epsilon]\big), \\ (\textit{distributed}): &\quad \big([e^* \pm \epsilon]\,,\, 0\,,\, e_{pi} + e_{ci} \in [e^* \pm \epsilon]\big). \end{aligned} \tag{A.12}$$

To show this, exclude out-of-equilibrium profiles:

a) Obviously, *no* investment profile can be an ORE where three or more players invest nothing.

b) There are three possible ORE where two players invest nothing:

$$\begin{aligned}(i): &\quad \big([e^* \pm \epsilon]\,,\, 0\,,\, 0\,,\, [e^* \pm \epsilon]\big),\\ (ii): &\quad \big([e^* \pm \epsilon]\,,\, 0\,,\, [e^* \pm \epsilon]\,,\, 0\big),\\ (iii): &\quad \big(0\,,\, [e^* \pm \epsilon]\,,\, [e^* \pm \epsilon]\,,\, 0\big).\end{aligned}$$

Profiles (i) and (ii) are contained in the classes of ORE described above. In profile (iii), the *sum* of $c1$'s and $c2$'s investments must, by Definition 3, be weakly smaller than $e^* + \epsilon$. Hence, profile (iii) is *not* an ORE when $2(e^* - \epsilon) > e^* + \epsilon$ and thus when $\epsilon < e^*/3$.

c) There are two ORE where one player invests nothing:

$$\begin{aligned}(iv): &\quad \big([e^* \pm \epsilon]\,,\, 0\,,\, e_{c_2} + e_{p_2} \in [e^* \pm \epsilon]\big),\\ (v): &\quad \big(0\,,\, [e^* \pm \epsilon]\,,\, e_{c_2} + e_{p_2} \in [e^* \pm \epsilon]\big).\end{aligned}$$

Profile (iv) is contained in the classes of ORE described above. Profile (v) is *not* an equilibrium when for player c_2:

$$\max\{e_{c_2}\} = e^* + \epsilon \quad < \quad \min\Big\{\sum_{i \in N} e_i\Big\} = 2(e^* - \epsilon)$$

and hence when $e^* + \epsilon < 2(e^* - \epsilon) \;\Leftrightarrow\; \epsilon < e^*/3$.

d) When *all* players make a positive investment, it follows from the best-response conditions of the end players $p1$ and $p2$ that

$$e_{p_i} + e_{c_i} \in [e^* \pm \epsilon]\,. \tag{A.13}$$

At the same time, the best response of a middle player requires that

$$e_{p_i} + e_{c_i} + e_{c_j} \in [e^* \pm \epsilon]\,. \tag{A.14}$$

Combining (A.13) and (A.14), it follows that

$$e_{p_i} \;\; \geq \;\; e^* - \epsilon - e_{c_i} \;\geq\; e^* - \epsilon - (e^* + \epsilon - e_{c_j} - e_{p_i}) \;\Leftrightarrow\; e_{c_j} \leq 2\epsilon$$

Hence, we get $0 < e_{c_i} \leq 2\epsilon$. Using (A.13) again, we moreover get $e^* - 3\epsilon \leq e_{p_i} < e^* + \epsilon$ and, thus, we arrive at a profile that is contained in the classes of ORE described above.

Refined ORE: When $\tau_c \in T_c$ for both center players $c \in C$, $\tau_p \in T_p$ for both peripheral players $p \in P$, and $\epsilon < e^*/5$, then

$$\pi_{c_i}(e) \geq \pi_{p_i}(e) \quad \forall\, i \in \{1,2\}\,. \tag{A.15}$$

To see this, suppose, to the contrary, that $\pi_{c_1}(e) < \pi_{p_1}(e)$ (or $\pi_{c_2}(e) < \pi_{p_2}(e)$ or both). Then, we must have a *distributed* profile with

$$\big(e_{p_2} = [e^* \pm \epsilon]\,,\, e_{c_2} = 0\,,\, e_{p_1} + e_{c_1} \in [e^* \pm \epsilon]\big)$$

because $\epsilon < e^*/5$ implies that in an *end-specialized* profile it holds $e_{ci} < e_{pi}$ and thus $\pi_{c_i}(e) > \pi_{p_i}(e)$. In particular, for a distributed profile to arise in ORE, the first-order conditions for the center player c_1 and the periphery player p_1 need to be satisfied, while at the same time, it must be $\pi_{c_1}(e) < \pi_{c_2}(e)$.[33] Hence, we require

$$\begin{aligned}
(i) \quad \frac{\partial U_{c_1}(e)}{\partial e_{c_1}} &= (b'_{c_1} - c)\big(1 - \rho_{c_1}\frac{|R^+_{c_1}|}{|R_{c_1}|} - \sigma_{c_1}\frac{|R^-_{c_1}|}{|R_{c_1}|}\big) + \frac{\sigma_{c_1}}{2}\big(b'_{p_1} + b'_{c_2}\big) \;=\; 0 \\
(ii) \quad \frac{\partial U_{p_1}(e)}{\partial e_{p_1}} &= (b'_{p_1} - c)\big(1 - \rho_{p_1}\frac{|R^+_{p_1}|}{|R_{p_1}|} - \sigma_{p_1}\frac{|R^-_{p_1}|}{|R_{p_1}|}\big) + \rho_{p_1} b'_{c_1} \;\leq\; 0\,.
\end{aligned}$$

However, it follows from the same argument as made for the star network that condition (i) implies $\partial U_{p_1}(e)/\partial e_{p_1} > 0$. A contradiction to the necessary equilibrium condition (ii). In an ORE on the line network, payoff ranking (A.15) must therefore apply.

Refined ORE with limited preference strength: Payoff ranking (A.15) even translates into an investment ranking when $\epsilon < e^*/5$. To see this, note that, by (A.12), it holds $e_{c_i} < e_{p_i}$ for both $i \in \{1,2\}$ in an *end-specialized* equilibrium. Moreover, as condition (A.15) implies that $\pi_{c_i}(e) \geq \pi_{p_i}(e)$, we also get $e_{ci} \leq e_{pi}$ for both $i \in \{1,2\}$ in a *distributed* equilibrium. Thus, we have $\pi_{c_i}(e) \geq \pi_{p_i}(e)$ and $e_{p_i} \geq e_{c_i}$ for $i \in \{1,2\}$.

A.2.3 Dyad and complete network

ORE set: It immediately follows from Definition 3 that $\sum_{i \in N} e_i \in [e^* \pm \epsilon]$ must hold.

Refined ORE on dyad: We show that $\tau_i \in T_c \cap T_p$ for both players i, then it must be

$$e_i = e_j = e \in \big[\frac{e^* \pm \epsilon}{n}\big]\,. \tag{A.16}$$

To see this, note that utility in the dyad can be written as

$$U_i(e) \;=\; b\big(e_i + e_j\big) - ce_i + \rho_i |N_i^+|(e_i - e_j)c + \sigma_i |N_i^-|(e_i - e_j)c\,,$$

where $|N_i^+| = 1$ and $|N_i^-| = 0$ iff $\pi_i(e) > \pi_j(e) \;\Leftrightarrow\; e_i < e_j$. Suppose now that, contrary to (A.16), $e_i > e_j \geq 0$. For this to be an ORE, we require

$$\begin{aligned}
(i) \quad \frac{\partial U_i(e)}{\partial e_i} &= b' - c + \sigma_i c \;=\; 0 \\
(ii) \quad \frac{\partial U_j(e)}{\partial e_j} &= b' - c + \rho_j c \;\leq\; 0\,.
\end{aligned}$$

However, since $\rho_j \geq 0 \geq \sigma_i$ (where at least one inequality is strict since $\tau_i, \tau_j \in T_c \cap T_p$), conditions (i) and (ii) cannot be satisfied simultaneously. Hence, in an ORE, $e_i = e_j = e$ needs to hold.

[33] It must be $\pi_{c_1}(e) < \pi_{c_2}(e)$ because in a distributed profile, it is

$$\pi_{c_2}(e) \geq b(e_{c_1} + e^* - \epsilon)$$

and

$$\pi_{c_1}(e) = b(e_{c_1} + e_{p_1}) - ce_{c_1}\,.$$

Moreover, in a distributed profile, $\pi_{c_1}(e) < \pi_{p_1}(e)$ implies that $e_{c_1} > e_{p_1}$. Thus, suppose to the contrary that $\pi_{c_1}(e) \geq \pi_{c_2}(e)$. Then $e_{c_1} > e_{p_1} > e^* - \epsilon$ must hold. This is however incompatible with $e_{p_1} + e_{c_1} \in [e^* \pm \epsilon]$ when $\epsilon < e^*/3$.

Refined ORE on complete network: Suppose that $\tau_i \in T_c \cap T_p$ for all players i. Suppose moreover that ρ_i and ρ_j, respectively σ_i and σ_j, are sufficiently close together for all $i, j \in N$. Then, an ORE must entail the equal split in (A.16).

To show this, note that utility in the complete network can be written as

$$U_i(e) = b\Big(\sum_{i\in N} e_i\Big) - ce_i + \frac{\rho_i}{3}\sum_{j\in N_i^+}(e_i - e_j)c + \frac{\sigma_i}{3}\sum_{j\in N_i^-}(e_i - e_j)c\,.$$

Suppose now that, contrary to the statement, there are some players i and j with $e_i < e_j$. For this to be an ORE, it must hold for player i (j) with the lowest (highest) investment, for some small $h > 0$, and any $e_i' \in (e_i + h,\, e_i)$ and $e_j' \in (e_j - h,\, e_j)$ that

$$\begin{aligned}(i) \quad \frac{\partial U_i(e_i', e_{-i})}{\partial e_i} &= b' - c + \rho_i \frac{|N_i^+|}{3}c + \sigma_i \frac{|N_i^-|}{3}c \;\le\; 0\\ (ii) \quad \frac{\partial U_j(e_j', e_{-j})}{\partial e_j} &= b' - c + \rho_j \frac{|N_j^+|}{3}c + \sigma_j \frac{|N_j^-|}{3}c \;\ge\; 0\,.\end{aligned}$$

These two conditions cannot be met simultaneously, however, when ρ_i and ρ_j, respectively σ_i and σ_j, are sufficiently close together because for (i) and (ii) to be satisfied we need that

$$|N_i^+|\rho_i + |N_i^-|\sigma_i \;\le\; |N_j^+|\rho_j + |N_j^-|\sigma_j\,. \qquad \text{(A.17)}$$

And since $|N_i^+| \ge |N_j^+| + 1$ and $|N_i^-| \le |N_j^-| - 1$, (A.17) requires

$$\rho_i - \sigma_i \;\le\; |N_j^+|(\rho_j - \rho_i) + |N_j^-|(\sigma_j - \sigma_i)\,. \qquad \text{(A.18)}$$

Note now that $\tau_i \in T_c \cap T_p$ implies $\rho_i - \sigma_i > 0$. This however means that (A.18) cannot be met by any $i \in N$ when $\rho_j - \rho_i \le x$ and $\sigma_j - \sigma_i \le y$ for all $i, j \in N$ and some small $x, y > 0$. We thus arrive at a contradiction between the two necessary equilibrium conditions (i) and (ii). In an ORE, it must therefore be $e_i = e_j = e$.

A.2.4 Circle

ORE set: Suppose that $\epsilon < e^*/5$. Then, the ORE set on the circle resembles the Nash equilibrium set from the original game, that is, an ORE entails either a *specialized* or a *fully distributed* investment profile.

To show this, fix the sequence of players in the order i, j, k, l. Suppose first that $e_m > 0$ for all $m \in N$ (*fully distributed*). Based on Definition 3, every e_m must lie inside an interval $3\underline{e} \le e_m \le \bar{e}$, where

$$\underline{e} \,+\, 2\bar{e} \;=\; e^* - \epsilon \quad \text{and} \quad \bar{e} \,+\, 2\,\underline{e} \;=\; e^* + \epsilon\,.$$

Solving these equations and simplifying gives

$$e_m \in \Big[\frac{e^*}{3} \pm \epsilon\Big] \quad \text{for all } m \in N\,.$$

Next, suppose that $e_i = 0$ for some player i (*specialized*). It follows that i's neighbors, j and l, must make a positive investment because suppose, to the contrary, that $e_j = 0$ (or $e_l = 0$, or both are equal to zero). Then, $e_k > 0$ since otherwise $e_i + e_j + e_k = 0$. In fact, we need $e_k \ge e^* - \epsilon$ and $e_l \ge e^* - \epsilon$ for this to be a best-response profile for i and j. This however leads to a contradiction to the best-response condition of player k because when $e_l \ge e^* - \epsilon$ player k invests at most 2ϵ. Yet, this is at odds with $e_k \ge e^* - \epsilon$ when $\epsilon < e^*/3$. Thus, when $e_i = 0$ then it must be $e_j > 0$ and $e_l > 0$.

In fact, $e_i = 0$, $e_j > 0$, and $e_l > 0$ implies that $e_k = 0$ because suppose, to the contrary, $e_k > 0$. As the total investments received by players j, k, and l must satisfy

$$e_j + e_k \in [e^* \pm \epsilon] \,, \quad e_j + e_k + e_l \in [e^* \pm \epsilon] \,, \quad \text{and} \quad e_k + e_l \in [e^* \pm \epsilon]$$

respectively, it follows that $e_j \leq 2\epsilon$ and $e_l \leq 2\epsilon$. This however means that the total investment received by player i is no larger than 4ϵ. And when $\epsilon < e^*/5$, then $e_j + e_l \leq 4\epsilon < e^* - \epsilon$. A contradiction to $e_i = 0$. Thus, in an ORE on the circle, it must be $e_k = 0$ when $e_i = 0$. In particular, together with the equilibrium conditions for j and l, we get $(0\,,\, [e^* \pm \epsilon]\,,\, 0\,,\, [e^* \pm \epsilon])$.

A.2.5 General networks and incomplete information

In this section, we generalize our basic model from the main text to allow for incomplete information regarding the social preference types of the other players. Moreover, we provide the missing proof of Proposition 2 for a general network structure.

To incorporate incomplete information, suppose that the exact preference type τ_i of each player is privately known only to that individual. The exact types of the other players remain unknown. Suppose, however, that each player possesses a vague impression of the other players' preferences, possibly gained through prior encounters. Formally, let $\tau = (\tau_i)_{i \in N}$ represent one potential type constellation in $\Omega = T_1 \times ... \times T_n$, where the type sets T_i are potentially heterogeneous. Then, our assumption is that the probability function $p(\tau) : \Omega \rightarrow (0, 1)$ is common knowledge.

In line with our basic setting, each player's utility depends on her relative standing among the players in her reference group also in this setting as well. We assume that a player compares her own expected payoff with that of her peers. Formally, let $\tau_{-i} = (\tau_j)_{j \neq i}$ denote one potential type constellation for all other players $j \neq i$, and let $e_{-i} = (e_{\tau_j})_{\tau_{-i} \in \Omega_{-i}}$ denote the profile of investments for all possible types of $j \neq i$. The expected utility of a type-τ_i of player i at investments (e_{τ_i}, e_{-i}) is given by

$$\begin{aligned} \mathbb{E}_{\tau_{-i}}[U_i|\tau_i] \;\; &= \;\; \mathbb{E}_{\tau_{-i}}[\pi_i|\tau_i] \; + \; \frac{\sigma_{\tau_i}}{|R_i|} \sum_{j \in R_i^-} \left(\mathbb{E}_{\tau_{-i}}[\pi_j|\tau_i] - \mathbb{E}_{\tau_{-i}}[\pi_i|\tau_i] \right) \\ &\quad + \; \frac{\rho_{\tau_i}}{|R_i|} \sum_{j \in R_i^+} \left(\mathbb{E}_{\tau_{-i}}[\pi_j|\tau_i] - \mathbb{E}_{\tau_{-i}}[\pi_i|\tau_i] \right) , \end{aligned} \tag{A.19}$$

where R_i^- (R_i^+) denote the subsets of players in i's reference group who earn more (less) in expectation, with expected payoffs given by $\mathbb{E}_{\tau_{-i}}[\pi_i|\tau_i] = \sum_{\tau_{-i} \in \Omega_{-i}} p(\tau_{-i}|\tau_i) b(e_{\tau_i} + \sum_{k \in N_i} e_{\tau_k}) - ce_{\tau_i}$ and $\mathbb{E}_{\tau_{-i}}[\pi_j|\tau_i] = \sum_{\tau_{-i} \in \Omega_{-i}} p(\tau_{-i}|\tau_i) b(e_{\tau_j} + \sum_{k \in N_j} e_{\tau_k}) - ce_{\tau_j}$.

The following result, which generalizes Proposition 2 from the main text, can be verified in this extended setting:

Proposition 3. *Consider two players i and j in a nested neighborhood of a network g such that all their types have compatible social preferences (i.e., $T_i \subset T_c$ and $T_j \subset T_p$). In an ORE, it must then hold for at least one $\tau_i \in T_i$ and $\tau_j \in T_j$ that*

$$\mathbb{E}_{\tau_{-i}}[\pi_i|\tau_i] \geq \min_{k \in N_i} \{\mathbb{E}_{\tau_{-i}}[\pi_k|\tau_i]\} \qquad \mathit{OR} \qquad \mathbb{E}_{\tau_{-j}}[\pi_j|\tau_j] \leq \max_{l \in N_j} \{\mathbb{E}_{\tau_{-j}}[\pi_l|\tau_j]\} \,.$$

Proof. Suppose that, contrary to the statement, all types of player i earn strictly less in expectation than all $k \in N_i$ and all types of player j earn strictly more in expectation than all $l \in N_j$. One immediate implication is that $\mathbb{E}_{\tau_{-i}}[\pi_i|\tau_i] < \mathbb{E}_{\tau_{-i}}[\pi_j|\tau_i] \; \forall \, \tau_i \in T_i$. Because player j's neighborhood is nested in player i's, we moreover have for all $\tau \in \Omega$ that

$$e_{\tau_i} + \sum_{k \in N_i} e_{\tau_k} \;\; \geq \;\; e_{\tau_j} + \sum_{l \in N_j} e_{\tau_l} \,. \tag{A.20}$$

In combination, $e_{\tau_i} > 0 \;\forall\, \tau_i \in T_i$ thus needs to hold because i has access to more investments than j in any $\tau_{-i} \in \Omega_{-i}$, while at the same time i earns less in expectation.

The first-order conditions for all possible types $\tau_i \in T_i$ of player i and all possible $\tau_j \in T_j$ of player j thus become[34]

$$\begin{aligned}
(i) \quad \frac{\partial \mathbb{E}_{\tau_{-i}}[U_i|\tau_i]}{\partial e_{\tau_i}} \;&=\; \sum_{\tau_{-i} \in \Omega_{-i}} p(\tau_{-i}|\tau_i) \Bigg[\Big(b'\big(e_{\tau_i} + \sum_{k \in N_i} e_{\tau_k}\big) - c \Big) \Big(1 - \frac{|R^-_{\tau_i}|}{|R_{\tau_i}|}\rho_{\tau_i} - \frac{|R^-_{\tau_i}|}{|R_{\tau_i}|}\sigma_{\tau_i} \Big) \\
&\qquad + \frac{\sigma_{\tau_i}}{|R_{\tau_i}|} \sum_{k \in N_i} b'\big(e_{\tau_k} + \sum_{m \in N_k} e_{\tau_m}\big) \Bigg] \;=\; 0 \qquad \text{(A.21)} \\
(ii) \quad \frac{\partial \mathbb{E}_{\tau_{-j}}[U_j|\tau_j]}{\partial e_{\tau_j}} \;&=\; \sum_{\tau_{-j} \in \Omega_{-j}} p(\tau_{-j}|\tau_j) \Bigg[\Big(b'\big(e_{\tau_j} + \sum_{l \in N_j} e_{\tau_l}\big) - c \Big) \Big(1 - \frac{|R^-_{\tau_j}|}{|R_{\tau_j}|}\rho_{\tau_j} - \frac{|R^-_{\tau_j}|}{|R_{\tau_j}|}\sigma_{\tau_j} \Big) \\
&\qquad + \frac{\rho_{\tau_j}}{|R_j|} \sum_{l \in N_j} b'\big(e_{\tau_l} + \sum_{m \in N_l} e_{\tau_m}\big) \Bigg] \;\leq\; 0 \,.
\end{aligned}$$

Because all $\tau_i \in T_i$ and $\tau_j \in T_j$ have compatible social preferences, it is $\sigma_{\tau_i} \leq 0$ and $\rho_{\tau_j} \geq 0$ with at least one inequality being strict. For condition (i) to be satisfied for all $\tau_i \in T_i$, we thus need that

$$\sum_{\tau_{-i} \in \Omega_{-i}} p(\tau_{-i}|\tau_i)\, b'\big(e_{\tau_i} + \sum_{k \in N_i} e_{\tau_k}\big) \;\; \geq \;\; c \qquad \forall\, \tau_i \in T_i \,.$$

Summing up over all τ_i, this gives $\sum_{\tau \in \Omega} p(\tau)\, b'\big(e_{\tau_i} + \sum_{k \in N_i} e_{\tau_k}\big) \geq c$, or equivalently

$$\sum_{\tau_j \in T_j} p(\tau_j) \sum_{\tau_{-j} \in \Omega_{-j}} p(\tau_{-j}|\tau_j)\, b'\big(e_{\tau_i} + \sum_{k \in N_i} e_{\tau_k}\big) \;\; \geq \;\; c \,. \tag{A.22}$$

Because i nests the neighborhood of j (see (A.20)) and because $b(\cdot)$ is strictly concave, we additionally have

$$b'\big(e_{\tau_j} + \sum_{l \in N_j} e_{\tau_l}\big) \;\; \geq \;\; b'\big(e_{\tau_i} + \sum_{k \in N_i} e_{\tau_k}\big) \qquad \forall\, \tau \in \Omega \,. \tag{A.23}$$

In combination, (A.22) and (A.23) imply

$$\sum_{\tau_{-j} \in \Omega_{-j}} p(\tau_{-j}|\tau_j)\, b'\big(e_{\tau_i} + \sum_{k \in N_i} e_{\tau_k}\big) \geq c$$

for at least one $\tau_j \in T_j$ because otherwise the weighted average on the left-hand side of (A.22) could not be greater than c. Yet, together with the parameter conditions for preference compatibility, this means that the first-order condition in (A.21) is violated for at least one τ_j. In an ORE, payoffs must therefore be ordered as stated in the proposition. ■

At least two aspects of Proposition 3 are noteworthy: Firstly, for social preferences to result

[34]Here, we have implicitly assumed that $\mathbb{E}_{\tau_{-i}}[\pi_i(e_{\tau_i}, e_{-i})|\tau_i] \neq \mathbb{E}_{\tau_{-i}}[\pi_k(e_{\tau_i}, e_{-i})|\tau_i]$ for all $k \in R_i$ and that the same holds for player j and her peers. Nevertheless, because $U_i(e_{\tau_i}, e_{-i})$ is continuous, very similar first-order conditions must hold for all e'_{τ_i} and e'_{τ_j} in some small open neighborhoods around e_{τ_i} and e_{τ_j}.

in the predicted payoff ranking, it is imperative that *all* potential types of players i and j have compatible social preferences (i.e., $T_i \subset T_c$ and $T_j \subset T_p$). Otherwise, there could be a type of player j in T_j who is unwilling to contribute to the public good in case that player i's investment falls short of e^*. As a consequence, player i could not afford to lower his investment below e^*, even though the "real" type of player j is willing to fill the gap. Secondly, the assumption that the type sets of i and j are common knowledge is essential as well. Otherwise, a player j of the correct type might mistakenly believe that player i is in need, or player i might wrongly believe that j is not willing to contribute, etc. In other words, equilibrium refinement through social preferences not only requires a compatible preference combination but also a common understanding of this.

A.3 Measuring social preference strengths

Here, we establish a result to map a pair of social preference parameters, (ρ_i, σ_i), into an upper bound $\hat{\epsilon}_i$ for a player's true preference strength ϵ_i, which is valid for all the two- and four-player networks in our experiment.

Lemma 1. *Consider a player with utility function* (2) *and a quadratic payoff function* (1) *who occupies a position in one of the seven networks of Figure 1. An upper bound $\hat{\epsilon}_i$ for the player's true social preference strength ϵ_i is given by:*

- *for i in a nested position (e.g., periphery position in the star, core, d-box, or line, duo position in the core, or position in the dyad or complete network):*

$$\textit{altruists and social-welfare types}: \ \frac{\rho_i c e^*}{b'(0) - c}$$

$$\textit{inequity-averse types}: \ \max\Big\{\frac{-\sigma_i |R_i| c e^*}{(|R_i| - \rho_i(|R_i| - |N_i|))(b'(0) - c)} \ ; \ \frac{\rho_i c e^*}{b'(0) - c}\Big\}$$

$$\textit{competitive and spiteful types}: \ \frac{-\sigma_i c e^*}{b'(0) - c}$$

- *for i in a non-nested position (e.g., center position in the star, core-periphery, d-box, or line, or position in the circle):*

$$\textit{altruists and social-welfare types}: \ \frac{\rho_i b'\big(e^*/|N_i|\big) e^*}{b'(0) - c}$$

$$\textit{inequity-averse types}: \ \max\Big\{\frac{-\sigma_i((|N_i| - 1) b'(0) - c) e^*}{(|N_i| - \sigma_i(|N_i| - 1) - \rho_i \frac{(|R_i| - |N_i|)\,|N_i|}{|R_i|})(b'(0) - c)} \ ; \ \frac{\rho_i b'\big(e^*/|N_i|\big) e^*}{b'(0) - c}\Big\}$$

$$\textit{competitive and spiteful types}: \ \frac{-\sigma_i((|N_i| - 1) b'(0) - c) e^*}{(|N_i| - \sigma_i(|N_i| - 1))(b'(0) - c)} \, .$$

Proof. Our aim is to determine, for a given (ρ_i, σ_i) and a given network position i, an upper bound $\hat{\epsilon}_i$ for the difference between that player's best-response investment, $f_i(\tau_i, e_{-i})$, and a payoff-maximizing best response, $f_i(e_{-i})$, for all possible e_{-i}. More concretely, we aim to determine an e_i that constrains the deviation-maximizing best response in the following way:

$$\hat{\epsilon}_i \ \equiv \ \big|e_i - f_i(e_{-i})\big| \quad \geq \quad \epsilon_i \ \equiv \ \max\Big\{\big|f_i(\tau_i, e_{-i}) - f_i(e_{-i})\big| \ : \ \forall \, e_{-i} \in [0, \bar{e}]^{n-1}\Big\}$$

and that satisfies $(|\rho_i|, |\sigma_i|) < (|\rho_i'|, |\sigma_i'|) \Rightarrow \hat{e}_i(|\rho_i|, |\sigma_i|) < \hat{e}_i(|\rho_i'|, |\sigma_i'|)$.

Nevertheless, as utility function $U_i(e)$ is not differentiable at investments where $\pi_i(e) = \pi_j(e)$ for some $j \in R_i$, we need to make some case distinctions.

(I) Deviation-maximizing interior solutions: Suppose first that the deviation-maximizing $f_i(\tau_i, e_{-i})$ is such that $\pi_i(f_i(\tau_i, e_{-i}), e_{-i}) \neq \pi_j(f_i(\tau_i, e_{-i}), e_{-i})$ for all $j \in R_i$. Then, $f_i(\tau_i, e_{-i})$ needs to satisfy the first-order condition

$$\begin{aligned}\frac{\partial U_i}{\partial e_i} &= \Big(b'\big(f_i(\tau_i, e_{-i}) + \sum_{j \in N_i} e_j\big) - c\Big)\Big(1 - \rho_i \frac{|R_i^+|}{|R_i|} - \sigma_i \frac{|R_i^-|}{|R_i|}\Big) \\ &+ \frac{\sigma_i}{|R_i|} \sum_{j \in N_i^-} b'\big(f_i(\tau_i, e_{-i}) + e_j + \sum_{l \in N_j \setminus \{i\}} e_l\big) \\ &+ \frac{\rho_i}{|R_i|} \sum_{j \in N_i^+} b'\big(f_i(\tau_i, e_{-i}) + e_j + \sum_{l \in N_j \setminus \{i\}} e_l\big) \leq 0\,,\end{aligned} \tag{A.24}$$

where N_i^+ (N_i^-) denotes the set of neighbors with $\pi_i > (<)\, \pi_j$, and R_i^+ (R_i^-) the set of peers with $\pi_i > (<)\, \pi_j$. The corresponding condition for a payoff-maximizing investment $f_i(e_{-i})$ is

$$b'\big(f_i(e_{-i}) + \sum_{j \in N_i} e_j\big) - c \leq 0\,. \tag{A.25}$$

In the first step, we determine an upper bound for a *positive* deviation, $f_i(\tau_i, e_{-i}) > f_i(e_{-i})$, before we proceed to a lower bound for a *negative* deviation, $f_i(\tau_i, e_{-i}) < f_i(e_{-i})$.

(IA) Positive deviations: By definition of a positive deviation, it must be $f_i(\tau_i, e_{-i}) > 0$ so that the condition in (A.24) must be satisfied with equality and, moreover, $b'\big(f_i(\tau_i, e_{-i}) + \sum_{j \in N_i} e_j\big) - c < 0$ in the first line of (A.24).

Hence, to establish an upper bound for them, set $R_i^+ = R_i$ and $R_i^- = \emptyset$ in the first line of (A.24). Moreover, set $N_i^+ = R_i = N_i$ and $N_i^- = \emptyset$ in the second and third lines of (A.24). Because $\rho_i \geq \sigma_i$, this results in an increase in the terms in lines 1–3 and, consequentially, in $f_i(\tau_i, e_{-i})$, while leaving the condition in (A.25) and, by extension, the value for $f_i(e_{-i})$ unaffected.

Our upper bound e_i for $f_i(\tau_i, e_{-i})$ thus satisfies[35]

$$\begin{aligned}&\big(b'\big(e_i + \sum_{j \in N_i} e_j\big) - c\big)\big(1 - \rho_i\big) \\ &\qquad + \frac{\rho_i}{|N_i|} \sum_{j \in N_i} b'\big(e_i + e_j + \sum_{l \in N_j \setminus \{i\}} e_l\big) = 0\,.\end{aligned} \tag{A.26}$$

This immediately implies that $e_i - f_i(e_{-i}) > 0$ if and only if $\rho_i > 0$. Yet, to be able to continue from here, we need to make some additional case distinctions.

[35]Obviously, we ignore at this point the constraints on e_i and e_{-i} that are necessary for $\pi_i(e) > \pi_j(e)$ for all $j \neq i$. For this reason, e_i is just an upper bound for a best-response investment, but it may not be supported as a best response itself. However, e_i may in fact be a best-response investment for a certain player type in a certain network position.

(IA1) Positive deviations in nested positions: When i is in a nested network position, i.e., $N_i \cup \{i\} \subseteq N_j \cup \{j\}$ for all $j \in N_i$, equation (A.26) simplifies to

$$\begin{aligned}&\big(b'\big(e_i + \sum_{j \in N_i} e_j\big) - c\big)\big(1 - \rho_i\big) \\ &\qquad + \frac{\rho_i}{|N_i|} \sum_{j \in N_i} b'\big(e_i + \sum_{j \in N_i} e_j + \sum_{l \in N_j \setminus (N_i \cup \{i\})} e_l\big) \;=\; 0\,.\end{aligned} \tag{A.27}$$

Because we have $b''(e) = b''(e')$ for all $e, e' \in [0, \bar{e}]$, the total derivative of (A.27) gives for any player l who is not a neighbor of i (i.e., $l \in N_j \setminus (N_i \cup \{i\})$):

$$\frac{de_i}{de_l} \;\leq\; -\frac{\rho_i}{|N_i|} \;<\; 0\,.$$

Hence, to maximize $e_i - f_i(e_{-i})$, set $e_l = 0$. As a result, (A.27) becomes

$$\begin{aligned}& b'\big(e_i + \sum_{j \in N_i} e_j\big) \;-\; (1 - \rho_i)c \;=\; 0 \\ \Leftrightarrow \quad & e_i + \sum_{j \in N_i} e_j \;=\; (b')^{-1}\big((1 - \rho_i)c\big)\,.\end{aligned} \tag{A.28}$$

Now, because e_i and $\sum_{j \in N_i} e_j$ are perfect strategic substitutes in both (A.28) and (A.25) and because $(b')^{-1}\big((1 - \rho_i)c\big) > e^*$, decrease $\sum_{j \in N_i} e_j$ from an initial high level down to the point where the first-order condition (A.25) becomes just binding. We then get $f_i(e_{-i}) = 0$ and $\sum_{j \in N_i} e_j = e^*$ and, thus, get an upper bound of

$$\hat{e}_i \;=\; e_i - f_i(e_{-i}) \;=\; (b')^{-1}\big((1 - \rho_i)c\big) - e^*.$$

When we finally leverage the quadratic nature of function $b(\cdot)$, (A.28) can be written as $e_i + \sum_{j \in N_i} e_j = (b'(0) - (1 - \rho_i)c)/|b''|$ and (A.25) as $e^* = (b'(0) - c)/|b''|$. So, we finally get

$$\hat{e}_i \;=\; \rho_i \frac{c}{b'(0) - c} e^*\,. \tag{A.29}$$

It is important to note that this bound (along with all bounds to come) is even the smallest possible upper bound because $e_i = \hat{e}_i$ represents the best response on $\sum_{j \in N_i} e_j = e^*$ of, for instance, an altruistic player in a complete network. Note, moreover, that our assumption $|b''| > (2b'(0) - c)/\bar{e}$ ensures that $\hat{e}_i + e^* < \bar{e}$.

(IA2) Positive deviations in non-nested positions: Suppose next that i's neighborhood is *not* nested in the neighborhoods of all players in i's neighborhood (i.e., $N_i \cup \{i\} \not\subseteq N_j \cup \{j\}$ for some $j \in N_i$). Starting from equation (A.26) again, the total derivative gives in this case

$$\frac{de_i}{de_l} \;\leq\; -\frac{\rho_i}{|N_i|} \;<\; 0$$

for any $l \in N_j \setminus \{i\}$. Hence, for a maximal positive deviation, set $e_l = 0$. The total derivative, furthermore, gives for any $j \in N_i$

$$\frac{de_i}{de_j} \;=\; -\big(1 - \rho_i + \frac{x\,\rho_i}{|N_i|}\big) \;\geq\; -1 \;=\; \frac{df_i(e_{-i})}{de_j}\,,$$

where $x \in \{1, .., |N_i|\}$ depending on how often player j is herself a neighbor of other $k \in N_i \setminus \{j\}$. Hence, because $df_i(e_{-i})/de_j$ is the total derivative of the first-order condition (A.25) for a payoff maximizer, decrease $\sum_{j\in N_i} e_j$ from an initial high level down to the point where the first-order condition of a payoff maximizer is just satisfied with equality, that is, where $\sum_{j\in N_i} e_j = e^*$ and $f_i(e_{-i}) = 0$.

Now, as $b'(e) > b'(e')$ for $e < e'$ and as the term in line two of (A.26) increases in $b'(e_i + e_j + \sum_{l\in N_j\setminus\{i\}} e_l)$, a maximal positive deviation is attained in a network position i, where none of i's neighbors are neighbors themselves (e.g., the star center). Our upper bound $\hat{e}_i = e_i - f_i(e_{-i}) = e_i$ thus satisfies

$$\begin{aligned}
& \big(b'\big(\hat{e}_i + e^*\big) - c\big)\big(1-\rho_i\big) + \frac{\rho_i}{|N_i|}\sum_{j\in N_i} b'\big(\hat{e}_i + e_j\big) \;=\; 0 \\
\Leftrightarrow\quad & \big(b'\big(\hat{e}_i + e^*\big) - c\big)\big(1-\rho_i\big) + \rho_i b'\big(\hat{e}_i + e^*/|N_i|\big) \;=\; 0 \\
\Leftrightarrow\quad & \hat{e}_i \;=\; \rho_i \frac{b'\big(e^*/|N_i|\big)}{b'(0) - c} e^* \,. \qquad\qquad \text{(A.30)}
\end{aligned}$$

where, in lines 2 and 3, we made use of the quadratic nature of $b(\cdot)$.

(IB) Negative deviations: Start from equations (A.24) and (A.25), again. Note first that since $\rho_i \geq \sigma_i$ and $b' > 0$, it must be $\sigma_i < 0$ and there must be at least one $j \in N_i^-$ for player i to deviate downwards from a payoff-maximizing best response.

Now, rewrite (A.24) as

$$\begin{aligned}
& \Big(b'\big(f_i(\tau_i, e_{-i}) + \sum_{j\in N_i} e_j\big) \;-\; c\Big)\Big(1 - \rho_i\frac{|R_i^+| - |N_i^+|}{|R_i|} - \sigma_i\frac{|R_i^-| - |N_i^-|}{|R_i|}\Big) \\
+\;\; & \frac{\sigma_i}{|R_i|}\sum_{j\in N_i^-}\Big(b'\big(f_i(\tau_i, e_{-i}) + e_j + \sum_{l\in N_j\setminus\{i\}} e_l\big) \;-\; b'\big(f_i(\tau_i, e_{-i}) + \sum_{j\in N_i} e_j\big) \;+\; c\Big) \\
+\;\; & \frac{\rho_i}{|R_i|}\sum_{j\in N_i^+}\Big(b'\big(f_i(\tau_i, e_{-i}) + e_j + \sum_{l\in N_j\setminus\{i\}} e_l\big) \;-\; b'\big(f_i(\tau_i, e_{-i}) + \sum_{j\in N_i} e_j\big) \;+\; c\Big) \\
\leq\;\; & 0\,. \qquad\qquad \text{(A.31)}
\end{aligned}$$

Hence, to establish our lower bound e_i for $f_i(\tau_i, e_{-i})$, the expressions in lines 1–3 need to be minimized, while leaving condition (A.25) for a payoff maximizer unaffected. To achieve this, set $|R_i^-| = |N_i^-|$ in line 1 because $\rho_i \geq \sigma_i$ and $b'(f_i(\tau_i, e_{-i}) + \sum_{j\in N_i} e_j) > c$. To proceed from here, we need to make some further case distinctions.

(IB1) Negative deviations when i is linked to everyone: When i is linked to every other player, it is $|R_i^+| = |N_i^+|$ in the first line of (A.31). Moreover, the expressions in parentheses in lines 2 and 3 are strictly positive because $f_i(\tau_i, e_{-i}) + e_j + \sum_{l\in N_j\setminus\{i\}} e_l \leq f_i(\tau_i, e_{-i}) + \sum_{j\in N_i} e_j$ and thus $b'(f_i(\tau_i, e_{-i}) + e_j + \sum_{l\in N_j\setminus\{i\}} e_l) \geq b'(f_i(\tau_i, e_{-i}) + \sum_{j\in N_i} e_j)$. Therefore, to minimize $f_i(\tau_i, e_{-i})$, set $N_i^+ = \emptyset$ and $N_i^- = N_i = R_i$ in lines 2 and 3 because $\rho_i \geq \sigma_i$.

Suppose, now, that i is linked to every other player because i is in the dyad or complete network. Based on the above steps, (A.31) simplifies to

$$b'\big(\sum_{i\in N} e_i\big) - \big(1 - \sigma_i\big)c \;\;\leq\;\; 0\,. \qquad\qquad \text{(A.32)}$$

Because e_i and $\sum_{j\in_i} e_j$ are perfect strategic substitutes in both (A.32) and (A.26), i.e., $de_i/(d\sum_{j\in_i} e_j) = df_i(e_{-i})/(d\sum_{j\in_i} e_j) = -1$, decrease $\sum_{j\in N_i} e_j$ from an initial high level down to the point where condition (A.32) is just satisfied with equality, that is, where $e_i = 0$. Our lower bound $\hat{e}_i$ is thus given by $\hat{\epsilon}_i = f_i(e_{-i}) = e^* - \sum_{j\in N_i} e_j$, where, by (A.32),

$$\sum_{j\in N_i} e_j \;=\; (b')^{-1}\big((1-\sigma_i)c\big)\,.$$

When we finally make use of the quadratic nature of $b(\cdot)$, we can write

$$\hat{\epsilon}_i \;=\; \frac{-\sigma_i c}{b'(0)-c}e^*\,,$$

Suppose, next, that i is linked to every other player because i resides in the center position of the star, core, or d-box. Inequality (A.31) then becomes

$$\big(b'(\sum_{i\in N} e_i) - c\big)\big(1-\sigma_i\big) \;+\; \frac{\sigma_i}{n-1}\sum_{j\in N\setminus\{i\}} b'\big(e_i+e_j\big) \;\le\; 0\,. \tag{A.33}$$

Moreover, its total derivative with respect to e_j (when (A.33) is satisfied with equality) gives for any $j \in N_i$

$$\frac{de_i}{de_j} \;=\; -(1-\sigma_i+\frac{x\sigma_i}{n-1}) \;\le\; \frac{df_i(e_{-i})}{de_j} \;=\; -1\,,$$

where $x \in \{1,..,n-1\}$, depending on how often j is a neighbor of other $k \in N_i\setminus\{j\}$. Hence, to obtain a maximal negative deviation, decrease $\sum_{j\in N_i} e_j$ from an initial high level down to the point where condition (A.33) is just satisfied with equality, that is, where $e_i = 0$.

More concretely, because $b'(e) > b'(e')$ for $e < e'$ and since (A.33) is declining in $b'(e_i+e_j)$, a maximal negative deviation is obtained in the star center position where none of i's neighbors are neighbors themselves. Our lower bound $\hat{\epsilon}_i$ is then given by $\hat{\epsilon}_i = f_i(e_{-i}) = e^* - \sum_{j\in N_i} e_j$, where

$$\big(b'(\sum_{j\in N_i} e_j) - c\big)\big(1-\sigma_i\big) + \frac{\sigma_i}{n-1}\sum_{j\in N_i} b'\big(e_j\big) \;=\; 0\,.$$

Making use of the quadratic function nature of $b(\cdot)$, again, we can write

$$\begin{aligned}
&\big(b'(\sum_{j\in N_i} e_j) - c\big)\big(1-\sigma_i\big) + \sigma_i b'\big(\frac{\sum_{j\in N_i} e_j}{n-1}\big) \;=\; 0\\
\Leftrightarrow\quad &\sum_{j\in N_i} e_j \;=\; \frac{b'(0)-c(1-\sigma_i)}{(1-\sigma_i+\frac{\sigma_i}{n-1})(b'(0)-c)}e^*\,.
\end{aligned}$$

So, we finally get

$$\hat{\epsilon}_i \;=\; -\sigma_i\frac{(n-2)b'(0)-c}{(n-1-\sigma_i(n-2))(b'(0)-c)}e^*\,.$$

(IB2) Negative deviations when i has a single neighbor: Start from (A.31), again. Because for a negative deviation (i.e., $f_i(\tau_i, e_{-i}) < f_i(e_{-i})$) we require $|N_i^-| > 0$, we immediately get $|N_i^-| = |N_i| = 1$ and $|N_i^+| = 0$ in lines 2 and 3 of (A.31). Moreover, for our maximal negative deviation, in line 1, set $|R_i^+| = |R_i| - |N_i|$ if $\rho_i > 0$, and $|R_i^+| = |N_i^+|$ if $\rho_i \le 0$.

Now, because the total derivative of (A.31) with respect to e_l, $l \in N_j\setminus\{i\}$ (when (A.31) is satisfied with equality) is

$$\frac{de_i}{de_l} \;=\; \begin{cases} -\frac{\sigma_i}{|R_i|-\rho_i(|R_i|-|N_i|)} & \text{if } \rho_i > 0 \\ -\frac{\sigma_i}{|R_i|} & \text{otherwise} \end{cases} ,$$

we have, $de_i/de_l > 0$. Hence, to minimize e_i, set $e_l = 0$ for all $l \in N_j \backslash \{i\}$.

As a result, the term in parentheses in line 2 of (A.31) becomes strictly positive. Hence, set $|R_i| = |N_i^-| = 1$ in this line, so that inequality (A.31) reduces to

$$\begin{cases} \big(b'(e_i+e_j) - c\big)\big(1-\rho_i \frac{|R_i|-|N_i|}{|R_i|}\big) + \sigma_i c \;\leq\; 0 & \text{if } \rho_i > 0 \\ b'(e_i+e_j) - c(1-\sigma_i) \;\leq\; 0 & \text{otherwise} \end{cases} . \tag{A.34}$$

Finally, because e_i and e_j are perfect strategic substitutes in both (A.34) and (A.26), decrease $\sum_{j\in N_i} e_j$ from an initial high level down to the point where (A.34) is just satisfied with equality, that is, where $e_i = 0$. Our lower bound $\hat{\epsilon}_i$ is, thus, given by $\hat{\epsilon}_i = f_i(e_{-i}) = e^* - e_j$, where

$$e_j \;=\; \begin{cases} (b')^{-1}\big((1-\frac{|R_i|\sigma_i}{|R_i|-\rho_i(|R_i|-|N_i|)})c\big) & \text{if } \rho_i > 0 \\ (b')^{-1}\big((1-\sigma_i)c\big) & \text{otherwise} \end{cases} .$$

For a quadratic function $b(\cdot)$, we then get

$$\hat{\epsilon}_i \;=\; \begin{cases} \frac{-|R_i|\sigma_i c}{(|R_i|-\rho_i(|R_i|-|N_i|))(b'(0)-c)} e^* & \text{if } \rho_i > 0 \\ \frac{-\sigma_i c}{b'(0)-c} e^* & \text{otherwise} \end{cases} . \tag{A.35}$$

(IB3) Negative deviations when i has two neighbors: Start from (A.31), again. Suppose first that $\sigma_i \leq \rho_i \leq 0$. Then, to minimize the term in line 1 of (A.31), set $|R_i^+| = |N_i^+|$. Moreover, for the same reasons as in (IB2), set $e_l = 0$ for $l \in N_j \backslash (N_i \cup \{i\})$ and $N_i^- = N_i = R_i$ in lines 2 and 3. Therefore, we have $\sum_{l\in N_j\backslash\{i\}} e_l = \sum_{l\in N_j\cap N_i} e_l$ so that (A.31) simplifies to

$$\begin{aligned} & b'(e_i + \sum_{j\in N_i} e_j) \;-\; c \\ & + \frac{\sigma_i}{|N_i|} \sum_{j\in N_i} \Big(b'(e_i + e_j + \sum_{l\in N_j\cap N_i} e_l) \;-\; b'(e_i + \sum_{j\in N_i} e_j) \;+\; c \Big) \;\leq\; 0\,. \end{aligned} \tag{A.36}$$

Suppose next that $\rho_i > 0 > \sigma_i$. To minimize the term in line 1 of (A.31), now, set $|R_i^+| = |R_i| - |N_i^-|$. Regarding the terms in lines 2 and 3, note that in all network positions with two neighbors (i.e., the line center, the circle, the d-box periphery, or the duo positions of the core), i's neighbors have no more than one neighbor l, $l \in N_j \backslash (N_i \cup \{i\})$, of their own. Remember, moreover that we require $|N_i^-| > 0$. Because we have $|\rho_i| \leq |\sigma_i|$ when $\rho_i > 0 > \sigma_i$ and because $b(\cdot)$ is a quadratic function, set $e_l = 0$, $N_i^+ = \emptyset$, and $N_i^- = N_i$. Therefore, we get similar to (A.36):

$$\begin{aligned} & \Big(b'(e_i + \sum_{j\in N_i} e_j) \;-\; c \Big)\Big(1-\rho_i \frac{|R_i|-|N_i|}{|R_i|}\Big) \\ & + \frac{\sigma_i}{|N_i|} \sum_{j\in N_i} \Big(b'(e_i + e_j + \sum_{l\in N_j\cap N_i} e_l) \;-\; b'(e_i + \sum_{j\in N_i} e_j) \;+\; c \Big) \;\leq\; 0\,. \end{aligned} \tag{A.37}$$

When we now assume that i has two neighbors because i resides in the d-box periphery or a core duo position, we get $e_j + \sum_{l\in N_j\cap N_i} e_l = \sum_{j\in N_i} e_j$. Hence, (A.36) and (A.37) simplify to the condition (A.34) for a player with a single neighbor. Hence, our lower bound $\hat{\epsilon}_i$ is given by

(A.35).

Assume, next, that i is in a line center or circle position. Then, $N_j \cap N_i = \emptyset$. Moreover, (A.36) and (A.37) become

$$\begin{cases} \big(b'\big(e_i + \sum_{j\in N_i} e_j\big) - c\big)\big(1 - \sigma_i - \rho_i \frac{|R_i|-|N_i|}{|R_i|}\big) & \\ \quad + \frac{\sigma_i}{|N_i|}\sum_{j\in N_i} b'\big(e_i + e_j\big) \;\leq\; 0 & \text{if } \rho_i > 0 \\ \big(b'\big(e_i + \sum_{j\in N_i} e_j\big) - c\big)\big(1 - \sigma_i\big) & \\ \quad + \frac{\sigma_i}{|N_i|}\sum_{j\in N_i} b'\big(e_i + e_j\big) \;\leq\; 0 & \text{if } \rho_i \leq 0 \end{cases} \tag{A.38}$$

Now, because

$$\frac{de_i}{de_j} \;<\; \frac{df_i(e_{-i})}{de_j} \;=\; -1\,,$$

decrease $\sum_{j\in N_i} e_j$ from an initial high level down to the point where condition (A.38) is just satisfied with equality, that is, where $e_i = 0$. Our lower bound $\hat{e}_i$ is then given by $\hat{e}_i = f_i(e_{-i}) = e^* - \sum_{j\in N_i} e_j$, where $\sum_{j\in N_i} e_j$ solves

$$\begin{cases} \big(b'(\sum_{j\in N_i} e_j) - c\big)\big(1 - \sigma_i - \rho_i \frac{|R_i|-|N_i|}{|R_i|}\big) + \sigma_i b'\big(\frac{\sum_{j\in N_i} e_j}{|N_i|}\big) \;=\; 0 & \text{if } \rho_i > 0 \\ \big(b'(\sum_{j\in N_i} e_j) - c\big)\big(1 - \sigma_i\big) + \sigma_i b'\big(\frac{\sum_{j\in N_i} e_j}{|N_i|}\big) \;=\; 0 & \text{if } \rho_i \leq 0 \end{cases}$$

When we now make use of the quadratic function nature of $b(\cdot)$, we get

$$\hat{e}_i \;=\; \begin{cases} \frac{-\sigma_i((|N_i|-1)b'(0)-c)}{(|N_i|-\sigma_i(|N_i|-1)-\rho_i\frac{(|R_i|-|N_i|)|N_i|}{|R_i|})(b'(0)-c)} e^* & \text{if } \rho_i > 0 \\ \frac{-\sigma_i((|N_i|-1)b'(0)-c)}{(|N_i|-\sigma_i(|N_i|-1))(b'(0)-c)} e^* & \text{if } \rho_i \leq 0 \end{cases}.$$

(II) Deviation-maximizing corner solutions: Next, we establish upper bounds for a deviation-maximizing $f_i(\tau_i, e_{-i})$ in the cases where $f_i(\tau_i, e_{-i})$ involves at least one player j in i's reference group, such that $\pi_j(, e_{fi}(\tau_i, e_{-i}) - i) = \pi_i(f_i(\tau_i, e_{-i}), e_{-i})$.

We start with the case $f_i(\tau_i, e_{-i}) > f_i(e_{-i})$. Note that even though $U_i(\cdot)$ is not differentiable at $f_i(\tau_i, e_{-i})$, a best-response investment must still satisfy for some small $h > 0$ and all $e_i' \in (f_i(\tau_i, e_{-i}) - h, f_i(\tau_i, e_{-i}))$:

$$\frac{\partial U_i(e_i', e_{-i})}{\partial e_i} \;\;\geq\;\; 0\,. \tag{A.39}$$

Let us ignore the requirements on e_i' and e_{-i} for a moment that lead to $\pi_j(e_i', e_{-i}) < (>) \pi_i(e_i', e_{-i})$ and assume that $R_i^+(e_i') = R_i^+(f_i(\tau_i, e_{-i}))$ and $R_i^-(e_i') = R_i^-(f_i(\tau_i, e_{-i}))$ for all $e_i' \in (f_i(\tau_i, e_{-i}) - h, f_i(\tau_i, e_{-i}))$.[36] Then, the inequality in (A.39) suggests that our upper bound is given by the (weakly) larger e_i that satisfies the first-order condition (A.24) for an interior solution in (IA) with equality. In other words, an upper bound for a deviation-maximizing corner solution is just given by the upper bounds developed in (IA).

Next, consider the case of a negative deviation with $f_i(\tau_i, e_{-i}) < f_i(e_{-i})$. A best-response investment must then satisfy for some small $h > 0$ and all $e_i' \in (f_i(\tau_i, e_{-i}), f_i(\tau_i, e_{-i}) + h)$:

$$\frac{\partial U_i(e_i', e_{-i})}{\partial e_i} \leq 0\,. \tag{A.40}$$

That condition is, however, identical to condition (A.24) in (IB). A lower bound for a deviation-

[36] We implicitly made this same assumption at several places before.

Table 8: Frequency of refined equilibria

Network	Equilibrium type	Deviation from payoff-maximizing equilibrium: zero ($\chi = 0$)	moderate ($\chi < 3$)	any (any χ)
Dyad	equal split	32.1% (S,E,Q,rfd)	45.8% (rfd)	49.2% (rfd)
	other	8.8% (S,E)	33.0%	50.8%
Complete	equal split	0.8% (S,E,Q,rfd)	0.8% (rfd)	0.8% (rfd)
	other	20.8% (S,E)	62.5%	99.2%
Star	per-spec.	15.8% (S,Q,rfd)	33.3% (rfd)	62.5% (rfd)
	distr.: $\pi_c \geq \pi_j$	—	—	36.6% (rfd)
	cent-sp. or distr.	0% (E)	0.8%	0.8%
Circle	specialized	7.5% (S,E,rfd)	15.8% (rfd)	29.2% (rfd)
	distributed	3.3% (Q,rfd)	27.5% (rfd)	55.0% (rfd)
Core	per-spec.	17.5% (S,Q,rfd)	43.3% (rfd)	68.3% (rfd)
	distr.: $\pi_c \geq \pi_j$	—	—	31.7% (rfd)
	cent-sp. or distr.	0% (E)	0%	0%
D-box	per-spec.	8.3% (S,E,Q,rfd)	15.0% (rfd)	25.8% (rfd)
	distr.: $\pi_c \geq \pi_j$	—	1.7% (rfd)	64.2% (rfd)
	cent-sp. or distr.	0% (E)	9.2%	10.0%
Line	per-spec.	0.8% (S,Q,rfd)	40.1% (rfd)	49.2% (rfd)
	distr: $\pi_m \geq \pi_e$	8.3% (S,rfd)	13.3% (rfd)	16.7% (rfd)
	cent-sp. or distr.	1.6% (S,E)	8.3%	34.1%

NOTES: Percentages of (refined) Nash equilibrium profiles at the random round ends. Refined equilibria are: (Q) quantal response, (S) stable, (E) efficient, (rfd) refined other-regarding equilibria.

maximizing corner solution is just the lower bound of (IB). ■

B Experimental Appendix

B.1 Alternative refinement concepts

Here, we compare the predictive power of our refined ORE concept with that of several alternative equilibrium refinement concepts applied to the Bramoullé and Kranton (2007) game. Table 8 summarizes the predictions of the most relevant concepts:

- *Asymptotically stable* equilibria based on the idea that, in our continuous-time experiment, a best-response dynamic might lead back to a stable equilibrium following a single mistake by a player.
- *Efficient* equilibria rooted in the idea that subjects might utilize the time we give them to coordinate on an equilibrium maximizing group welfare.
- *Quantal response (logit) equilibria* (McKelvey and Palfrey, 1995) based on the idea that subjects play best responses to the fluctuating choices of their network neighbors.

As demonstrated in Table 8, particularly in Column 3 ($\chi = 0$), the alternative refinement concepts do not explain the experimental findings better than our preferred theory in any network structure. On the contrary, the *efficiency* concepts fares worse across all networks, either because it fails to refine the equilibrium set in certain networks or because it selects the "wrong" equilibria. The predictive power of efficiency is particularly low in the star and the core-periphery network, where it is efficient when the public good is entirely provided by the center player but where most investments are made in the peripheral positions (see Table 8).[37]

Asymptotic stability fares better than efficiency, especially in the star, core-periphery, and d-box. Nevertheless, it fails to predict the empirically highly relevant equal-split equilibria on the dyad, as all equilibrium profiles are asymptotically stable on this network.

Only the *quantal-response* concept comes close to our refined ORE predictions. As demonstrated by Rosenkranz and Weitzel (2012), the theory selects a unique Nash equilibrium profile on all the seven networks in our experiment when players make marginal decision errors. Moreover, the selected equilibria align with our refined ORE predictions in most of these networks. Yet, quantal-response theory tends to generate a too fine-grained selection for the circle network, where it predicts an equal split of twelve units as the only equilibrium outcome, even though a specialized equilibrium is even more prevalent in the data. Similarly, on the line network, quantal-response theory predicts a periphery-specialized equilibrium even though a partially distributed public good is more frequently observed.

B.2 Distribution of social preference types and strengths

Table 9 outlines the results of our classification of each subject's $(\hat{\rho}_i, \hat{\sigma}_i)$-pair into its revealed preference type and its revealed preference strength.

Table 9: Revealed preference types and strengths

	Preference strength						
		in nested positions		in center positions		in line middle and circle	
	any	moderate	marginal	moderate	marginal	moderate	marginal
Preference type	(any $\hat{\epsilon}_i$)	($\hat{\epsilon}_i < 3$)	($\hat{\epsilon}_i < 1$)	($\hat{\epsilon}_i < 3$)	($\hat{\epsilon}_i < 1$)	($\hat{\epsilon}_i < 3$)	($\hat{\epsilon}_i < 1$)
altruism	11.7%	11.7%	10.0%	9.2%	2.5%	10.0%	4.2%
social welfare	15.0%	15.0%	14.2%	11.7%	0.8%	10.0%	0.8%
inequity averse	29.2%	29.2%	5.8%	4.2%	0%	12.5%	0%
competitive	10.0%	10.0%	2.5%	2.5%	0%	5.8%	0%
spiteful	23.3%	15.8%	9.2%	10.0%	6.7%	15.0%	6.7%
payoff maximizer	4.2%	4.2%	4.2%	4.2%	4.2%	4.2%	4.2%
asocial	6.7%	6.7%	6.7%	1.7%	0%	1.7%	0%
	100.0%	100.0%	47.5%	41.7%	14.2%	62.5%	15.8%

NOTES: Categorization of estimated $(\hat{\sigma}_i, \hat{\rho}_i)$-pairs into revealed preference types and revealed preference strengths. Insignificant estimates (i.e., p-values ≥ 0.05) or estimates with $-0.05 \leq x \leq 0.05$ for $x \in \{\hat{\sigma}_i, \hat{\rho}_i\}$ are set to zero because a subject with such a small parameter would make a decision indistinguishable from a payoff maximizer in our experiment.

B.3 Measurement error in tests of Hypothesis 1

In this appendix, we present the outcomes of our sensitivity tests for Hypothesis 1, where we introduced measurement error in our social preference estimates.

The underlying assumption for all our tests is that the random assignment of subjects to groups has effectively worked in our experiment, rendering our preference compatibility indicator truly exogenous. Hence, without measurement error, a comparison between the shares of refined ORE for groups with compatible and incompatible social preferences yields

[37]This is not entirely surprising. As suggested by Charness et al. (2014), efficiency concerns are particularly powerful in games where equilibrium outcomes can be Pareto ranked. Such a ranking is, however, not possible in our game with strategic substitutes.

an unbiased and consistent estimator for the true effect of preference compatibility, denoted as $P(\text{ref ORE} \mid c) - P(\text{ref ORE} \mid i)$, where c stands for compatible and i for incompatible.

By contrast, with measurement error, misclassification of several subject groups into having the right or wrong preference combination is likely. To assess the resulting bias, let $P(c)$ and $P(i) = 1 - P(c)$ denote the likelihoods that a group truly has (in-)compatible social preferences for a certain network. Both these likelihoods depend on the network-specific requirements outlined in Sections 3.4.1 and 3.4.2. Moreover, let $P(\hat{i}|c)$ and $P(\hat{c}|i)$ denote the conditional likelihoods of a misclassification, which depend on the measurement error in our preference estimates. The shares of refined OREs played by groups with seemingly compatible or incompatible preferences were then measured in expectation:

$$\begin{aligned}
\mathbb{E}[r_{\hat{c}}] &= \frac{P(\text{ref ORE} \mid c)P(\hat{c}|c)P(c) + P(\text{ref ORE} \mid i)P(\hat{c}|i)P(i)}{P(\hat{c}|c)P(c) + P(\hat{c}|i)P(i)} \\
\mathbb{E}[r_{\hat{i}}] &= \frac{P(\text{ref ORE} \mid i)P(\hat{i}|i)P(i) + P(\text{ref ORE} \mid c)P(\hat{i}|c)P(c)}{P(\hat{i}|i)P(i) + P(\hat{i}|c)P(c)},
\end{aligned}$$

where $P(\hat{c}|c) = 1 - P(\hat{i}|c)$ and $P(\hat{i}|i) = 1 - P(\hat{c}|i)$. Thus, when our theory is correct, and $P(\text{ref ORE} \mid c) > P(\text{ref ORE} \mid i)$, then $r_{\hat{c}} - r_{\hat{i}}$ underestimates the true effect of preference compatibility (higher type II error). In contrast, if our theory is incorrect, and $P(\text{ref ORE} \mid c) = P(\text{ref ORE} \mid i)$, measurement error does not distort the estimated effect (same type I error).

For our sensitivity checks, we solved the above equation system for $P(\text{ref ORE} \mid c) - P(\text{ref ORE} \mid i)$. We then utilized the social preference estimates from Table 1 of Fehr and Charness (2023) to determine the likelihoods $P(c)$ and $P(i)$ for a typical WEIRD student population. Our key assumption here is that our own subject pool is a representative sample of this population.[38] Finally, we simulated the misclassification probabilities $P(\hat{i}|c)$ and $P(\hat{c}|i)$ based on various assumptions about the underlying measurement error at the individual level. In one specification, we simulated slight measurement error, assuming that an ill-measured preference type is only one type "to the right" from a subject's true preference type on the scale: altruist-social welfare-inequity averse-money maximizer-competitive-spiteful. This one-sided deviation is motivated by the fact that our own preference estimates suggest a more "competitive" subject pool than the typical WEIRD student population. For our second specification, we simulated a more significant measurement error, assuming that an ill-measured type is randomly drawn from the other five preference types. In both specifications, we additionally varied the misclassification probabilities p.

[38]The estimates reviewed in Table 1 of Fehr and Charness (2023) suggest a combined share of 40% altruists and social-welfare types, 10% inequity-averse types, 45% money maximizers, and 5% competitive and spiteful types. Using the estimates from Bruhin et al. (2019) in addition, we then parsed the first group into 15% altruists and 25% social-welfare types and the last group into 2.5% competitive types and 2.5% altruists.

Table 10: Measurement error in tests of Hypothesis 1

Error type	observed	neighbor		random		observed	neighbor		random	
Error probability	refined ORE	0.1	0.3	0.1	0.3	refined ORE	0.1	0.3	0.1	0.3
Groups with	Any preference strength (any $\hat{e}$)					Moderate preference strength ($\hat{e} < 3$)				
Star										
Compatible pref.	1.00	1.00	1.00	1.00	1.00	0.16	0.16	0.16	0.16	0.17
Incompatible pref.	0.96	0.96	0.95	0.95	0.93	0.07	0.07	0.06	0.06	0.02
Dyad										
Compatible pref.	0.61	0.65	0.79	0.64	0.79	0.58	0.63	0.78	0.61	0.78
Incompatible pref.	0.41	0.41	0.35	0.39	0.24	0.36	0.35	0.29	0.33	0.17
Line										
Compatible pref.	0.78	0.79	0.8	0.79	0.84	0.57	0.58	0.60	0.59	0.67
Incompatible pref.	0.70	0.70	0.7	0.70	0.69	0.43	0.43	0.43	0.43	0.41
Core-periphery										
Compatible pref.	1.00	1.00	1.00	1.00	1.00	0.35	0.35	0.35	0.35	0.35
Incompatible pref.	0.99	0.99	0.99	0.99	0.99	0.29	0.28	0.27	0.28	0.27
Avg. diff.	0.08	0.10	0.15	0.10	0.20	0.13	0.15	0.21	0.15	0.28

NOTES: Observed shares of refined ORE and estimated shares of refined ORE (corrected for measurement error) for the four asymmetric networks supporting Hypothesis 1. Shares are shown separately for groups with compatible and incompatible social preferences.

The results of our sensitivity tests are summarized in Table 10. Columns 2 and 7 reproduce the observed ORE shares ($r_{\hat{c}}$ and $r_{\hat{i}}$) for the four asymmetric networks that lent support to our Hypothesis 1, as already seen in Table 3. Columns 3–6 and 8–11 then present our estimated ORE shares, corrected for measurement error.

C Replication instructions

C.1 Experimental design

Our computerized experiment was programmed in z-tree 3.0 (Fischbacher, 2007) and took place at the Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University between June 9 and June 18, 2008. We used the ORSEE recruitment system (Greiner, 2015) to invite over 1,000 potential subjects for our study via email.

During the experiment, the participating students played a local public goods game on the seven networks illustrated in Figure 1. A total of eight experimental sessions, each lasting approximately one-and-a-half hours, were scheduled and successfully completed. On average, 15 students participated in each session, resulting in a total of 120 participants across eight sessions. No student attended more than one session.

A typical session encompassed seven treatments (networks) with the treatment-ordering detailed in Table 11. At the commencement of each session, participants received general instructions, as shown below. Following the instructions, they played the local public goods game on each of the seven networks, repeating the same treatment five times in a row. Each set of five repetitions, referred to as rounds, included one trial round and four payoff-relevant rounds. To ensure anonymity, all choices were made in a manner that precluded their association with individual participants after the rounds or at the end of the experiment.

Table 11: Order of treatments by session

Session	Ordering	Treatment						
		1	2	3	4	5	6	7
1	1	Dyad	Line	Star	Circle	Core	D-box	Complete
2	2	Complete	D-box	Core	Circle	Star	Line	Dyad
3	3	Dyad	Star	Line	Core	Circle	D-box	Complete
4	4	Complete	D-box	Circle	Core	Line	Star	Dyad
5	3	Dyad	Star	Line	Core	Circle	D-box	Complete
6	2	Complete	D-box	Core	Circle	Star	Line	Dyad
7	1	Dyad	Line	Star	Circle	Core	D-box	Complete
8	4	Complete	D-box	Circle	Core	Line	Star	Dyad

At the onset of each round, participants were randomly assigned to new groups, consisting of either one (in the dyad) or three other participants (in all other networks). Participants were visually represented as circles on their computer screens, with self-identification facilitated by color (see screenshot below for an illustration).

Every round followed the same structure and lasted between 30 and 90 seconds. Starting from zero investments, participants could freely adjust their investments by clicking on two buttons at the bottom of their screens. Full information about the momentary investments of all other participants was continuously provided and updated five times per second. Also, the resulting payoffs of all participants were continuously displayed on their screens. Nevertheless, the actual points earned in a round were solely determined by the momentary investments of the players at the random round end, where investments were frozen and payoffs were counted. These round ends were randomly determined by the computer through a draw from the uniform distribution on the interval $[30, 90]$.

Taking the seven treatments together, each participant took part in 35 rounds within 35 distinct groups, of which 28 were payoff-relevant. At the end of the experiment, the experimental points were converted into euros at a rate of 400 points = 1 Euro and discretely disbursed to the participants. In addition, participants received a 3 Euro show-up fee.

C.2 Experimental instructions

-Instructions-

Please read the following instructions carefully. These instructions state everything you need to know in order to participate in the experiment. If you have any questions, please raise your hand. One of the experimenters will approach you to answer your question. The rules are equal for all the participants.

You can earn money by means of earning points during the experiment. The number of points that you earn depends on your own choices and the choices of other participants. At the end of the experiment, the total number of points that you earn will be exchanged at an exchange rate of:

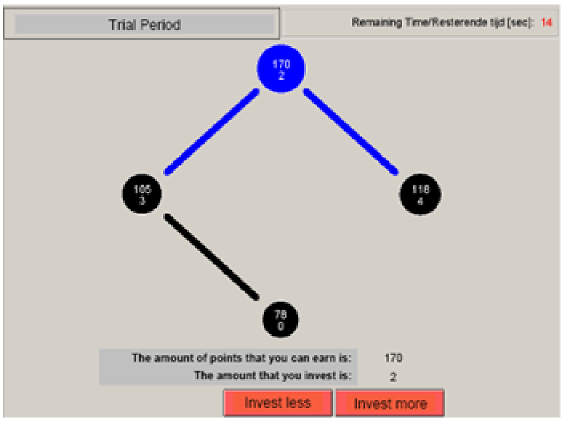
$$400 \text{ points} = 1 \text{ Euro}$$

The money you earn will be paid out in cash at the end of the experiment without other participants being able to see how much you earned. Further instructions on this will follow in due time. During the experiment, you are not allowed to communicate with other participants. Turn off your mobile phone and put it in your bag. Also, you may only use the functions on the screen that are necessary for the functioning of the experiment. Thank you very much.

-Overview of the Experiment-

The experiment consists of seven scenarios. Each scenario consists again of one trial round and four paid rounds (altogether 35 rounds, of which 28 are relevant for your earnings).
In all scenarios, you will be grouped with either one or three other randomly selected participants. At the beginning of each of the 35 rounds, the groups and the positions within the groups will be randomly changed. The participants that you are grouped within one round are very likely different participants from those you will be grouped within the next round. It will not be revealed with whom you were grouped at any moment during or after the experiment.
The participants in your group (of two or four players, depending on the scenario) will be shown as circles on the screen (see Figure 1). You are displayed as a **blue** circle, while the other participants are displayed as **black** circles. You are always connected to one or more other participants in your group. These other participants will be called your neighbors. These connections differ per scenario and are displayed as lines between the circles on the screen (see also Figure 1).
Each round lasts between 30 and 90 seconds. The end will be at an unknown and random moment in this time interval. During this time interval, you can earn points by producing know-how, but producing know-how also costs points. The points you receive in the end depend on your own investment in know-how and the investments of your neighbors.

Figure 1



By clicking on one of the two buttons at the bottom of the screen, you increase or decrease your investment in know-how. At the end of the round, you receive the amount of points that is shown on the screen at that moment in time. In other words, your final earnings only depend on the situation at the end of every round. Note that this end can be at any time between 30 and 90 seconds after the round is started, and that this moment is unknown to everybody. Also, different rounds will not last equally long.
The points you will receive can be seen as the top number in your blue circle. The points others will receive are indicated as the top number in the black circles of others. Next to this, the size of the circles changes with the points that you and the other participants will receive: a larger circle means that the particular participant receives more points. The bottom number in the circles indicates the amount invested in know-how by the participants in your group.
Remarks:

- It can occur that there is a time-lag between your click and the changes of the numbers on the screen. One click is enough to change your investment by one. A subsequent click will not be effective until the first click is effectuated.
- **Therefore wait until your investment in know-how is adapted before making further changes!**

-Your Earnings-

Now we explain how the number of points that you earn depends on the investments. Read this carefully. Do not worry if you find it difficult to grasp immediately. We also present an example with calculations below. Next to this, there is a trial round for each scenario to gain experience with how your investment affects your points.
In all scenarios, the points you receive at the end of each round depend in the same way on two factors:

1. **Every unit that you invest in know-how yourself will cost you 5 points.**
2. **You earn points for each unit that you invest yourself and for each unit that your neighbors invest.**

If you sum up all units of investment of yourself and your neighbors, the following table gives you the points that you earn from these investments:

Your investment plus your neighbors' investments	0	1	2	3	4	5	6	7	8	9	10
Points	0	28	54	78	100	120	138	154	168	180	190

Your investment plus your neighbors' investments	11	12	13	14	15	16	17	18	19	20	21
Points	198	204	208	210	211	212	213	214	215	216	217

The higher the total investments, the lower are the points earned from an additional unit of investment. Beyond an investment of 21, you earn one extra point for every additional unit invested by you or one of your neighbors.
Note: if your and your neighbors' investments add up to 12 or more, earnings increase by less than 5 points for each additional unit of investment.

-Example-

Suppose

1. you invest 2 units;
2. one of your neighbors invests 3 units and another neighbor invests 4 units.

Then you have to pay 2 times 5 = 10 points for your own investment. The investments that you profit from are your own plus your neighbors' investments: $2 + 3 + 4 = 9$. In the table, you can see that your earnings from this are 180 points. In total, this implies that you receive $180 - 10 = 170$ points if this would be the situation at the end of the round. Figure 1 shows this example as it would appear on the screen. The investment of the fourth participant in your group does not affect your earnings. In the trial round before each of the seven scenarios, you will have time to get used to how the points you will receive change with investments.

-Scenarios-

All rounds are basically the same. The only thing that changes between scenarios is whether you are in a group of two or four participants and how participants are connected to each other. Also, your own position randomly changes within scenarios and between rounds. We will notify you each time on the screen when a new scenario and trial round starts. At the top of the screen, you can also see when you are in a trial round (see top left in Figure 1). Paying rounds are just indicated by "ROUND" while trial rounds are indicated by "TRIAL ROUND".

-Questionnaire-

After the 35 rounds, you will be asked to fill in a questionnaire. Please take your time to fill in this questionnaire accurately. In the meantime, your earnings will be counted. Please remain seated until the payment has taken place.