Responsibility Sharing in a Public Bad Experiment[*]

Simin He[†] Xun Zhu[‡] Xinlu Zou[§]

October 13, 2025

Abstract

We propose a novel mechanism to mitigate the provision of public bads in large groups. In the baseline case with centralized punishment, players choose their neighbors and having more neighbors brings benefits. Each player then decides whether to provide a public bad, which yields personal benefits but imposes costs on the entire group. A small chance exists that a player providing the public bad will be detected and punished. In the responsibility-sharing mechanism, when a player is caught providing the public bad, their neighbors are also punished. Our theoretical analysis and experimental results strongly support the effectiveness of this mechanism in promoting Pareto optimal outcomes. In addition, using another experimental treatment, we determine that the mechanism is less effective in the absence of complete feedback information.

Keywords: Public bads, Cooperation, Punishment, Network formation, Laboratory experiment.

JEL codes: C72, C92, D82, D85.

[*] We thank Lata Gangadharan, Lingbo Huang, Theo Offerman, Dmitry Ryvkin, Ning Sun, Erte Xiao, Maoliang Ye, Jingjing Zhang, the audiences at Wuhan BEEF, Lyon ESA, and various seminars for comments and feedback. Simin He acknowledges the National Natural Science Foundation of China (grant 72473089) for financial support. Xun Zhu acknowledges the National Natural Science Foundation of China (grant 72503127) for financial support.

[†] School of Economics, Shanghai University of Finance and Economics, 777 Guoding Rd, 200433 Shanghai, China. E-mail: he.simin@mail.shufe.edu.cn

[‡] School of Digital Economics, Shanghai University of Finance and Economics, 777 Guoding Rd, 200433 Shanghai, China. E-mail: zhuxun@mail.shufe.edu.cn

[§] Corresponding author. School of Finance, Dongbei University of Finance and Economics, 217 Jianshan Street, 116023 Dalian, Liaoning, China. E-mail: zouxl0517@gmail.com

1 Introduction

Global public bads such as environmental pollution, traffic congestion, overfishing, and cigarette smoking present significant challenges to societies, resulting in severe social and economic consequences such as irreversible ecological damage, worsened health outcomes, and decreased productivity. While mitigating public bads requires cooperation among individuals, such cooperation has often proven difficult based on theoretical and experimental studies (Andreoni, 1988; Isaac and Walker, 1988a,b). Nevertheless, the literature has also identified a range of effective mechanisms for addressing this issue (for overviews, see Chen, 2008; Gächter and Herrmann, 2009; Chaudhuri, 2011).[1]

In a world where individuals belong to various social networks, one possible approach for sustaining cooperation is allowing people to choose their interactions (Cinyabuguma et al., 2005; Page et al., 2005; Rand et al., 2011; Charness and Yang, 2014; van Leeuwen et al., 2019). Cooperation often succeeds in such settings because many players adopt *link reciprocity* strategies, forming connections with cooperative individuals and avoiding noncooperative ones (Ebel and Bornholdt, 2002; Santos et al., 2006; Hanaki et al., 2007). These studies typically assume that public bads (or goods) are local, meaning individuals are unaffected by externalities outside their communities (Cornes and Sandler, 1996). However, this assumption does not hold in circumstances concerning global public bads such as greenhouse gas emissions, in which countries cannot isolate themselves from the effects of others' actions. In such cases, *link reciprocity* may fail, as excluding noncooperative players will not prevent external harm. This raises the question of whether network structures can still enhance cooperation in global public bad contexts.

This study explores mechanisms to deter the provision of global public bads in network structures. We introduce a novel responsibility-sharing mechanism, in which individuals form networks and only share responsibility with direct connections. The mechanism assumes that agents are motivated to connect with others, driven by intrinsic factors, such as regional free-trade agreements, or external incentives, such as government promotion opportunities based on the number of connections.[2]

[1] Studies on the public bads problem can be integrated into the framework of public goods game (Shitovitz and Spiegel, 2003).

[2] Exogenous sources can include information or technology sharing among interconnected individuals or companies (Bala and Goyal, 2000; Jackson and Watts, 2002; Goeree et al., 2009; Falk and Kosfeld, 2012; Jackson and Zenou, 2015).

We introduce a two-stage game. In the first stage, each player proposes links to others in the group, with a neighbor only forming if both parties propose it and having more neighbors brings benefits. In the second stage, players observe the network structure and decide whether to provide a public bad, which offers personal benefits but imposes costs on society. Crucially, the benefits of the public bad are less than the total costs it generates, creating a social dilemma. While abstaining from providing the public bad is socially optimal, providing it remains the dominant strategy for all players. We assume complete information, meaning players learn the choices made by others at the end of the game.

We introduce centralized punishment (CP) and responsibility-sharing (RS) mechanisms into the game. The CP mechanism reflects real-world systems, in which each player faces a small, exogenous probability of detection and subsequent punishment if caught providing the public bad. The RS mechanism extends the CP model by not only imposing punishment on the detected player, but also on their direct neighbors within the network. Our focus is investigating behavior and cooperation efficiency within the RS mechanism, with the CP mechanism serving as a benchmark to control for factors unrelated to RS.

Theoretically, both the CP and RS mechanisms can support full-cooperation and noncooperation outcomes as subgame-perfect equilibria in the infinitely repeated game. In the noncooperation equilibrium, all players provide the public bad. Under CP, players form a large network where everyone is connected, while the network is smaller under RS, with each player having one neighbor at most. In both the mechanisms, the full-cooperation equilibrium is efficient and yields higher payoffs for all players compared with the noncooperation equilibrium. This outcome is achieved when players form all links and refrain from providing the public bad. However, based on the basin of attraction theory (Dal Bó and Fréchette, 2011), we find that the full-cooperation equilibrium is more attractive in RS than in CP when the discount factor is sufficiently large. In the RS mechanism, the noncooperation equilibrium becomes less appealing due to the loss of network benefits, as a larger network cannot be supported in this equilibrium.

In our laboratory experiment, we use a between-subjects design to compare the behavior of the two mechanisms in CP and the RS treatments. In both treatments, participants are randomly assigned to eight-player groups and play the respective game in an infinitely repeated format, with random termination determined by a continuation probability of 0.97 after each period. The experimental results reveal a stark contrast between the two treatments. Nearly all groups in the RS treatment converge to full cooperation with complete networks, while no groups in the CP treatment achieve this outcome. Specifically, in the RS treatment, all participants, especially cooperative ones, are more likely to strategically form links with other cooperators than with noncooperators, drawing on reputations established from past behavior. This sorting behavior encourages noncooperative participants to switch to cooperation when they realize it is more profitable, resulting in full cooperation within the group. In contrast, participants' link decisions in the CP treatment do not depend on others' past behavior, resulting in a majority continuing to provide the public bad.

Moreover, assuming complete information among all players may be unrealistic. To address this, we introduce a private information (PI) treatment, which closely mirrors the RS treatment but with the key difference that players' public bad actions remain private. Players also have the option of revealing their actions at a cost. In the experiment, similar to the RS treatment, half of the PI groups achieve full cooperation, while the other half exhibit a pattern more like the CP treatment. Analyses reveal that successful cooperation in PI groups is driven by a higher disclosure rate, which establishes an environment that is more closely aligned with the RS treatment. Notably, participants rarely disclose their actions once full cooperation is achieved, suggesting that while information disclosure is crucial for fostering initial cooperation, it becomes less essential once cooperation is established.

Finally, while this is the first study to apply responsibility sharing to improve cooperation in a global public bad game, the concept has been widely employed in practical management, particularly in local game settings across fields such as finance, accounting, environmental protection, and import-export activities.[3] By extending these principles to the global context, the RS mechanism offers practical solutions to mitigate

[3] For instance, microfinance institutions in developing countries employ group-lending contracts, where loans are provided collectively to a group of low-income borrowers, with the entire group held liable if any member defaults on repayment (Morduch, 1999; Attanasio et al., 2014). Similarly, when banks identify a higher proportion of bad borrowers in a pool, they tighten lending standards, making it more difficult for all entrepreneurs in the pool to secure loans (Fishman et al., 2020; Fan, 2021). For more examples, see Oei (2017), Oei and Ring (2018) and The Pie News (2023).

global public bads, which are easily implementable when a centralized institution can detect and punish misbehavior, and when individuals have sufficient externally provided or naturally occurring incentives to form networks. For example, to combat corruption among government officials, authorities could implement policies that reward individuals with more connections and a willingness to share responsibility for corruption within their networks. Similarly, in international climate agreements, countries could receive enhanced trade benefits and developmental opportunities by forming more connections, contingent on substantial cooperation in sharing responsibility for potential violations.

2 Related Literature

This study is primarily related to three streams of literature. First, it contributes to enhancing the understanding of cooperative behavior within network structures. Previous research has predominantly focused on local public goods and bads, where individuals are unaffected by the impacts of these goods or bads beyond their immediate communities. For example, Cinyabuguma et al. (2005) find that in the voluntary contribution mechanism, groups exhibit nearly complete cooperation when players face potential exclusion based on majority voting. Similar results under different regrouping rules are observed by Page et al. (2005) and Charness and Yang (2014). In network structures, where players can autonomously choose their interaction partners, Rand et al. (2011) observe increased cooperation when players have the ability to frequently reconfigure links with others.[4] The success of cooperation in these studies is often attributed to the use of *link reciprocity*, which means players strategically form links with cooperative players and sever ties with noncooperative players (Ebel and Bornholdt, 2002; Santos et al., 2006; Hanaki et al., 2007; Shirado et al., 2013; Gallo and Yan, 2015; Riedl et al., 2021).[5]

In a closely related study, Riedl et al. (2016) investigate a two-stage game similar to

[4] He and Zou (2024) explore public goods provision within a network formation game, finding that cooperative behavior is highly sensitive to the costs of establishing links.

[5] The use of *link reciprocity* can be influenced by several factors. For example, van Leeuwen et al. (2019) find that players with central positions in a network are able to maintain connections with others even when their cooperation levels are relatively low. For more related findings, see Gächter and Thöni (2005), Ahn et al. (2009), Fowler and Christakis (2010), Brekke et al. (2011), Gracia-Lázaro et al. (2012), and Li et al. (2018).

ours, with the distinction that their second stage is a weakest-link coordination game. In their design, each player selects a single effort level that applies to all neighbors and payoffs are determined by the lowest effort level within their endogenously formed network. They find that players widely adopt *link reciprocity* to facilitate coordination in large groups with 24 players. Our study also relates to Riedl et al. (2016) in several noteworthy ways. First, we show that in endogenously formed networks, participants do not reach socially optimal outcomes in our global public bad game, in contrast to the findings of Riedl et al. (2016) for weakest-link games. A potential explanation is that attaining socially optimal outcomes in a public bad game requires players to choose dominated actions, whereas this is not the case in weakest-link games. Therefore, the public bad game environment is arguably more demanding. However, we identify a very effective RS mechanism through which socially optimal outcomes can be sustained in our setting. Finally, it is worth noting that the combination of responsibility sharing with probabilistic centralized punishment in a global public bad game establishes an environment that resembles a weakest-link setting, insofar as an individual's payoff may depend on the behavior of the least cooperative neighbors in the network.

Second, our study contributes to the literature on the role of punishment in fostering collective cooperation. Previous studies have shown that under decentralized punishment institutions —where individuals can punish others— cooperators frequently punish free-riders, even when doing so entails personal costs (Fehr and Gächter, 2000; Masclet et al., 2003; Nikiforakis and Normann, 2008; Nikiforakis, 2008; Xiao and Houser, 2011; Cason and Gangadharan, 2015; Boosey and Isaac, 2016).[6] While such behavior compels potential free-riders to cooperate in order to avoid punishment, the misuse of punishment rights may lead to antisocial punishment and reduce overall cooperation efficiency (Herrmann et al., 2008; Nikiforakis and Normann, 2008; Chaudhuri, 2011). Next, under CP institutions, where an authority imposes penalties, cooperation tends to improve more substantially

[6] Several studies explore the relationship between network structures, that determine who can punish whom and decentralized punishment (Carpenter et al., 2012; Leibbrandt et al., 2015; DeAngelo and Gee, 2020). The general finding is that complete networks tend to yield the most efficient outcomes. Moreover, Leibbrandt et al. (2015) find that network configurations has a more significant influence than punishment capacities on determining the level of public good provision. DeAngelo and Gee (2020) further demonstrate that when networks are endogenously formed, peer monitoring proves more effective than group monitoring.

when these institutions are endogenously chosen rather than exogenously imposed (Tyran and Feld, 2006; Bó et al., 2010; Putterman et al., 2011). This effect is particularly pronounced when exogenously imposed punishment institution lack sufficient deterrence, such as in case of low detection rates or weak punishment severity (Anderson and Stafford, 2003; Tyran and Feld, 2006; Bó et al., 2010; Andreoni and Gee, 2012; Kamijo et al., 2014). The RS mechanism proposed in this study enhances the deterrent power of CP without introducing additional management costs, provided that individuals are sufficiently motivated to form connections with others.

Finally, our study extends the concept of community enforcement, which has long been recognized as a central mechanism for sustaining social cooperation (Ostrom, 1990; Ostrom et al., 1992; Greif, 1989, 1993; Greif et al., 1994; Dixit, 2003). Previous literature primarily conceptualizes community enforcement as deterring noncooperative behavior through the threat of widespread cooperation breakdown across society (Kandori, 1992; Ellison, 1994; Takahashi, 2010; Wolitzky, 2013; Ali and Miller, 2014; Deb, 2020). In this framework, the victim of a defector typically adopts a grim trigger strategy wherein lacking a direct means of punishing the defector, they withdraw cooperation from all members of the community as an indirect form of punishment. This response spreads nooncooperation throughout the community and ultimately punishes the original defector. In contrast, the community enforcement induced by the RS mechanism in our study differs in two notable ways. First, punishment is confined to the defector's immediate network, ensuring that uninvolved parties outside this network are not subject to collateral punishment. Second, unlike traditional community enforcement, where the victim assumes the role of enforcer, the RS mechanism assigns this role to a centralized institution. This distinction enables more targeted punishment and reduces the risk of escalation into widespread noncooperation.[7] Finally, although the victim could, in principle, punish the entire community by withholding cooperation, our results demonstrate that the availability of targeted punishment under the RS mechanism renders such strategies rarely used.

[7] It can therefore be interpreted as a form of community enforcement through specialized enforcement mechanisms. Several studies have compared the relative roles of these two forms of enforcement in sustaining cooperation (Masten and Prüfer, 2014; Aldashev and Zanarone, 2017; Acemoglu and Wolitzky, 2020).

3 Theory

In this section, we present a model with costless links, which is consistent with the parameters used in the experiment. In Appendix A, we extend the model to a generalized version that allows for positive link costs.

3.1 Model

We construct a public bad network game. In the first stage, players endogenously form a network, where having more neighbors increases individual payoffs. In the second stage, players simultaneously decide whether to provide a public bad that imposes costs on all other players. In the baseline model, all players face a positive possibility of detection and those detected for choosing the public bad are punished. In the main model, we introduce a RS mechanism under which the neighbors of a detected provider are also punished.

3.1.1 Public bad network game

Eight players participate in the game, indexed as $N = \{1, 2, .., 8\}$. In the first stage, players simultaneously and independently choose who to propose links to. Player $i \in N$ chooses a set of proposals $\mathbf{g}_i = \{g_{i1}, g_{i2}, ..., g_{i8}\}$, where $g_{ij} = 1$ if i proposes a link to player j and $g_{ij} = 0$ otherwise. The induced network $\mathbf{g} = \{\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_8\}$ is a directed graph that represents the set of all proposals. Next, we adopt the two-sided link formation framework proposed by Jackson and Wolinsky (1996), where mutual consent is required to establish a neighbor relationship. In other words, players i and j become *neighbors* if and only if they propose a link to one another. Therefore, the closure of $\mathbf{g}$ is an undirected network that represents the set of neighbors, denoted by $\bar{\mathbf{g}} = \{\bar{g}_{ij} | \bar{g}_{ij} = g_{ij}g_{ji}, \forall i, j \in N\}$. Denote $N_i(\mathbf{g}) = \{j \in N | g_{ij}g_{ji} = 1, j \neq i\}$ as the set of player i's neighbors (equivalently, $N_i(\mathbf{g}) = \{j \in N | \bar{g}_{ij} = 1, j \neq i\}$), and let $|N_i(\mathbf{g})|$ denote the number of i's neighbors. Suppose the cost of each link proposal is c. For the purposes of our experiment, we assume $c = 0$, which simplifies participants' decision-making and aligns with a common assumption in the experimental literature on network games (Rand et al., 2011; Riedl et al., 2016; Teteryatnikova and Tremewan, 2020; Goyal et al., 2021).[8] In addition, we

[8] In the generalized model presented in Appendix A, we relax the assumption of $c = 0$ in two ways. (1) We introduce a positive cost for forming a neighbor through bilateral proposals, and (2) in Appendix A. 3, we examine the case of costly unilateral link proposals ($c > 0$).

introduce a scale effect within the network game, wherein each neighbor yields a benefit of 20. In other words, player i receives an additional benefit of $20|N_i(\mathbf{g})|$ by establishing $|N_i(\mathbf{g})|$ neighbors.

In the second stage, after observing the network $\mathbf{g}$ formed in the first stage, each player simultaneously chooses an action, $a_i \in \{x, y\}$. Let $\mathbf{a} = (a_1, a_2, ..., a_8)$ denote all players' action profiles. Choosing x represents not providing the public bad, while choosing y represents providing it. A player who chooses x obtains a benefit of 50 and imposes no externalities on others, while a player who chooses y earns 100 but imposes a cost of 15 on every other group member. Therefore, although y maximizes individual payoffs, x is socially efficient.

Player i's final payoff is determined by two components, the gain and the loss (denoted as $U_{i,gain}$ and $U_{i,loss}$, respectively). The gain is calculated as the sum of the benefits derived from the network and the action in the second stage, which is represented as follows:[9]

$$U_{i,gain} = 20|N_i(\mathbf{g})| + 50(100), \text{if } a_i = x(y). \tag{1}$$

The loss equals the total costs imposed by the other players who choose y in the group, which is represented as follows:

$$U_{i,loss} = 15 \sum_{k \in N, k \neq i} \mathbb{I}\{a_k = y\}. \tag{2}$$

Finally, player i's payoff is the difference between the two components, as follows:

$$U_i(\mathbf{g}, \mathbf{a}) = U_{i,gain} - U_{i,loss}. \tag{3}$$

A key assumption of the public bad network game is that each additional neighbor yields a fixed benefit of 20. This reflects network economies of scale, whereby players derive direct benefits from expanding their connections. For example, a larger pool of Uber drivers increases consumer adoption of the platform, subsequently improving individual drivers' earnings. Similarly, firms located in larger industrial parks benefit from reduced transportation costs by leveraging synergies within the industrial chain.[10]

[9] We assume linearity in combining the benefits arising from the scale effect of the network and the payoffs associated with actions x or y.

[10] Various studies employ different formulations of network economies of scale. For example, in some settings a player's payoff increases proportionally with neighborhood size (Riedl et al., 2016). Notably, we demonstrate that the equilibrium properties derived below remain robust, employing a setting akin to that of Riedl et al. (2016).

3.1.2 Baseline model: CP with random detection

The baseline model incorporates CP with random detection into the public bad network game. Specifically, following the second-stage action choices, a central institution randomly detects each player's actions with a fixed detection probability. If a player who chose action y is detected, they incur a fixed punishment, setting their gain component of the payoff to 0, while their loss component remains unaffected.[11] Conversely, if a player who chose y is not detected or if a player who chose x is detected, no punishment is imposed, and the gain and loss components of their payoffs remain unaffected. In summary, the gain component of player i's payoff in the baseline model can be described as follows:

$$U_{i,gain} = \begin{cases} 20|N_i(\mathbf{g})| + 50(100), & \text{if } i \text{ is not detected to choose } y \text{ and } a_i = x(y) \\ 0, & \text{if } i \text{ is detected to choose } y \end{cases} . \quad (4)$$

We assume that the detection process is exogenous and independent across players, with a fixed probability of 15%. This relatively low detection rate implies that choosing action x is not a dominant strategy.

3.1.3 Main model: RS mechanism

We extend the baseline model by introducing the RS mechanism into the main model. Under this mechanism, if player i is detected choosing action y, then player i is punished with a zero gain and all of their neighbors also incur the same punishment. In other words, the consequences of providing the public bad extend to player i's immediate network. Similarly, the RS mechanism's punishment only affects the gain component of the payoff, which is presented as follows:

$$U_{i,gain} = \begin{cases} 20|N_i(\mathbf{g})| + 50(100), & \text{if neither } i \text{ nor any player in } N_i(\mathbf{g}) \text{ is detected} \\ & \text{ to choose } y, \text{ and } a_i = x(y) \\ 0, & \text{if } i \text{ or at least one player in } N_i(\mathbf{g}) \text{ is detected} \\ & \text{to choose } y \end{cases} . \quad (5)$$

[11] An alternative approach to model the punishment payoff is to incorporate a fixed deduction that players receive when they are punished. For the sake of simplicity, we choose to assign a fixed status quo payoff of zero to players upon being punished.

3.2 Equilibrium analysis

This section begins by examining the equilibrium of the one-shot game for baseline and main models. Owing to the dynamic nature of the two-stage game, we adopt the subgame-perfect equilibrium as the solution concept.[12] We eliminate weakly dominated strategies as a refinement of multiple equilibria in both models. We then analyze the equilibrium of the infinitely repeated game. In Appendix A, we demonstrate that all the equilibrium properties presented in this section hold in the generalized model.[13]

3.2.1 Equilibria of the one-shot game

We next derive the equilibrium for the one-shot game of the baseline model described in Section 3.1.2, and the main model described in Section 3.1.3.[14] The one-shot game includes two stages, which we analyze using backward induction. Proposition 1 characterizes the equilibrium strategies in the second-stage subgame.

Proposition 1. *For baseline and main models, given any network* $\boldsymbol{g}$*, the unique equilibrium strategy for the second-stage subgame is* $a_i = y$ *for all* $i \in N$*.*

Proposition 1 shows that the unique equilibrium strategy of the second-stage subgame for any network formed in the first stage in baseline and main models is all players choosing

[12] Although the network formation process in our first-stage game follows Jackson and Wolinsky (1996), we do not adopt their solution concept of pairwise stable networks because this study focuses on an infinitely repeated game in which the primary interests lie in the dynamic strategies underlying the formation of neighbor relationships.

[13] The generalized model introduces positive link cost for each neighbor formed, and all equilibrium properties continue to hold when this cost falls below a certain threshold. A cost higher than this threshold reduces the equilibrium set of the one-shot game, producing equilibrium outcomes with fewer neighbors in the first stage, without affecting the dominant choice of y in the second stage. In the alternative setting of costly unilateral proposal discussed in Appendix A. 3, we demonstrate that the equilibrium properties of the one-shot game are similar to those in Propositions 2 and 3, except that unilateral proposals no longer exist in equilibrium. Moreover, while the "elimination of weakly dominated strategies" used in Section 3.2.2 is not useful in the one-shot game, the equilibria of the infinitely repeated game remain valid.

[14] We do not derive the equilibrium of the public bad network game without CP introduced in Section 3.1.1, as it is straightforward that proposing to all other players and providing the public bad are dominant strategies.

action y. As shown in Appendix B, this is because the marginal benefit of choosing y always exceeds the expected punishment given the low detection rate.

Next, we derive the equilibrium for the one-shot game. Proposition 2 presents the equilibrium properties of the baseline model.

Proposition 2. *For every subgame-perfect equilibrium of the baseline model,* $(\boldsymbol{g}, \boldsymbol{a})$*, the following properties hold for all* $i, j \in N$ *with* $i \neq j$*: (i)* $g_{ij} = g_{ji}$ *and (ii)* $a_i = y$*.*

The first condition of Proposition 2 imposes no constraints on the equilibrium network structure, with the exception of the requirement that all link proposals in the network should be bilateral. The rationale behind this condition is that if a player unilaterally proposes a link to another player, it is always profitable for the latter to form a bilateral link by proposing back, irrespective of the choices made by both players in the second stage. Examples of equilibrium networks include the *complete* network, wherein each player proposes links to all others, and the *empty* network, where no links exist.

Proposition 3 presents the equilibrium properties of the main model.

Proposition 3. *For every subgame-perfect equilibrium of the main model,* $(\boldsymbol{g}, \boldsymbol{a})$*, the following properties hold for all* $i, j \in N$ *with* $i \neq j$*: (i)* $|N_i(\boldsymbol{g})| \leq 1$*, (ii) if* $|N_i(\boldsymbol{g})| < 1$*,* $g_{ji} = 0$*, and (iii)* $a_i = y$*.*

The first condition of Proposition 3 states that, in equilibrium, each player has at most one neighbor. This restriction arises because having more neighbors increases the risk of punishment in the main model and all players choose action y in the second stage. Moreover, the expected payoff initially rises with an increased number of neighbors but eventually declines. Consequently, we find that the optimal number of neighbors to maximize the expected payoff is one. The second condition asserts that, in equilibrium, a player should not unilaterally propose a link to someone with fewer than one neighbor. Otherwise, the latter player has an incentive to propose back to receive a higher payoff. One example of an equilibrium network is the 1-*regular* network in which all players have exactly one neighbor. Specifically, we divide the eight players into four disconnected pairs and players within each pair are neighbors to one another. Moreover, the empty network can also be an equilibrium network in the main model.

3.2.2 Eliminating weakly dominated strategies

In the baseline and main models, the unique equilibrium in the second-stage subgame is that all players provide the public bad. However, various network structures can be sustained in the first-stage equilibrium. In the baseline model, players can form any network structure without unilateral proposals, while in the main model, players can have one neighbor at most.

Given the multiplicity of equilibria in the one-shot game in both models, we introduce a refinement that eliminates weakly dominated strategies. Imposing this refinement enables us to focus on the equilibria in which players propose more links to one another under zero link costs. To illustrate, consider the equilibrium strategy of proposing zero links and choosing action y in the baseline model. This strategy is weakly dominated by the strategy of proposing links to all other players and choosing y because the latter yields a strictly higher payoff for the player when at least one other player proposes a link back, yielding a benefit from establishing more neighbors. The payoff remains unchanged when other players do not propose links to the player, given the zero link cost. Proposition 4 illustrates that, by imposing this refinement in the baseline model, the only surviving equilibrium is characterized by players forming a complete network and choosing action y. Conversely, the equilibrium in the main model involves players forming a 1-regular network and choosing action y.

Proposition 4. *After eliminating weakly dominated strategies, the baseline model has a unique, undominated subgame-perfect equilibrium,* $(\boldsymbol{g}, \boldsymbol{a})$*, where for all* $i, j \in N$ *with* $i \neq j$*: (i)* $g_{ij} = 1$ *and (ii)* $a_i = y$*, and the undominated subgame-perfect equilibrium of the main model,* $(\boldsymbol{g}, \boldsymbol{a})$*, satisfies the following properties for all* $i \in N$*: (i)* $|N_i(\boldsymbol{g})| = 1$ *and (ii)* $a_i = y$*.*

3.2.3 Equilibria of the infinitely repeated game

In this section, we examine the infinitely repeated game of baseline model and main models with a discount factor δ and derive the corresponding equilibrium. First, we demonstrate that adopting the undominated equilibrium strategy of the one-shot game in each period constitutes a subgame-perfect equilibrium in the infinitely repeated game. We refer to these equilibria as *noncooperation*, as all players choose action y in these equilibria.

Proposition 5 (Noncooperation equilibrium). *The strategy in which, for every period and history, each player consistently plays the equilibrium strategy of the one-shot game in the baseline and the main model outlined in Proposition 4, constitutes a subgame-perfect equilibrium of the infinitely repeated game of each respective model.*

In addition to the noncooperation equilibria, we further confirm that a full-cooperation outcome in which all players form a complete network and choose action x in each period can be supported by a subgame-perfect equilibrium of the infinitely repeated game with the following grim trigger strategy.[15]

Definition 1. *The grim trigger strategy is defined as follows:*

(i) For baseline and main models, in each period from the first period, each player proposes links to all other players in the first stage and chooses x in the second stage;

(ii) In the baseline model, if any player deviates from (i) in period $t \geq 1$, all players propose links to all other players in the first stage and choose y in the second stage in every period after t;

(iii) In the main model, if any player deviates from (i) in period $t \geq 1$, all players propose a single link, forming a fixed 1-regular network in the first stage and choose y in the second stage in every period after t.

Proposition 6 (Full-cooperation equilibrium). *When $\delta \geq 0.13$, the grim trigger strategy constitutes a subgame-perfect equilibrium of the infinitely repeated game of the baseline model. When $\delta \geq 0.06$, the grim trigger strategy constitutes a subgame-perfect equilibrium of the infinitely repeated game of the main model.*

Proposition 6 indicates that full cooperation can be sustained in the infinitely repeated game if players revert to an equilibrium of the one-shot game that survives the elimination of weakly dominated strategies following any deviation, as outlined in Proposition 4. Specifically, players in the baseline model are threatened by the risk of loss resulting from all other players choosing action y, whereas players in the main model not only incur losses

[15] Other outcomes in which players choose action x but establish different networks may also be supported in the infinitely repeated game. This study focuses on a complete network as it is the most focal network due to its efficiency.

when other players choose y, but also face the risk of losing neighbors, as they form a 1-regular network and choose y as a consequence of deviation. Notably, the full-cooperation equilibrium exists for both models. The detailed proof is presented in Appendix B.

3.2.4 Comparing the likelihood of cooperation

Although we have demonstrated that both noncooperation and full cooperation can be equilibrium in the infinitely repeated games of baseline and main models, the question of which equilibrium is more likely to occur remains unanswered. To address this, we employ the basin of attraction criterion proposed by Dal Bó and Fréchette (2011). The authors argue that the decision of whether to adopt a cooperative strategy in an infinitely repeated cooperation game depends on various factors such as game parameters and beliefs regarding the probability of other opponents adopting the cooperative strategy. Their findings reveal a negative correlation between the probability of evolving toward cooperation and the size of the basin of attraction in repeated-game strategies.

By considering these insights, we compare the likelihood of full cooperation between the two models based on their basin of attraction of the noncooperation strategy. The expected payoff of the full-cooperation equilibrium is 190 for each player in every period, regardless of the model type. However, the noncooperation equilibrium yields a payoff of 134 in the baseline model, whereas it only yields a payoff of 16.7 in the main model. This discrepancy arises from the fact that players in the noncooperation equilibrium of the main model can, at best, establish links with only one other player in the group. In contrast, in the baseline model, players benefit from establishing links with more players without being affected by their choices regarding the public bad. Intuitively, the lower payoff associated with the noncooperation equilibrium renders the full-cooperation equilibrium relatively more appealing in the main model compared with the baseline model.

Formally, we focus on two equilibrium strategies of the infinitely repeated game, encompassing the grim trigger strategy employed in the full-cooperation equilibrium, and the noncooperation strategy where players repeatedly play a one-shot game equilibrium that survives the elimination of weakly dominated strategies, as outlined in Proposition 5.[16] We calculate the size of the noncooperation strategy's basin of attraction respectively

[16] Proposition 5 allows players to establish different neighbors across periods and to propose unilateral links in the noncooperation equilibrium. However, for calculation simplicity, we restrict attention to

for baseline and main models, which is the probability that a player assigns to each opponent using the grim trigger strategy such that the player is indifferent between playing the two strategies. A larger basin of attraction indicates that a player must have a stronger belief in the opponents' cooperation to opt for the cooperation strategy (grim trigger), indicating a lower likelihood for cooperation to emerge. As a result, we arrive at the following proposition.[17]

Proposition 7 (Basin of attraction)**.** *If $\delta \geq 0.55$, the size of the noncooperation strategy's basin of attraction in the main model is smaller than that of the baseline model.*

This proposition indicates that in our model, where the discount factor is assumed to be sufficiently high, full cooperation is more likely to arise in the main model compared with the baseline model.

4 Experimental Design, Procedures, and Hypotheses

4.1 Experimental design

In the experiment, we employ a between-subjects design to implement three treatments: the Centralized Punishment treatment (CP), the Responsibility-Sharing treatment (RS), and the Private Information treatment (PI). In the CP and RS treatments, the game closely follows the baseline and main models, respectively. Each participant's link proposals and actions in CP and RS treatments are public information that can be observed by all other participants in the group. Additionally, in the PI treatment, participants play the game based on the main model, but their actions in the second stage are private and undisclosed, except when they are detected for choosing the public bad action. However, participants can voluntarily disclose their actions to the group at an additional small cost.

The CP and RS treatments are designed to identify the effect of the RS mechanism. The PI treatment is introduced for two main reasons. First, it enables us to decompose the factors that may contribute to improving cooperation in the RS treatment. Second, it mirrors real-world settings more closely and therefore has greater empirical relevance.

noncooperation strategies in which players form a fixed 1-regular network over periods without unilateral proposals.

[17] See Appendix B for the detailed basin of attraction calculation and proof of Proposition 7.

In many contexts, whether an agent has engaged in harmful behavior (a public bad) is private information to others in society. Examples include illicit tax evasion, pollutant emissions, or food safety violations, which are often conducted discreetly, sometimes without the knowledge of close partners or collaborators. However, in many cases, agents can voluntarily disclose their actions to the public, albeit at a cost. For example, chemical companies may publish emissions data on their websites to demonstrate environmental responsibility and food manufacturers can display images of their production facilities as evidence of compliance with safety standards.

In all three treatments, we conduct the infinitely repeated game in the laboratory through a process known as random termination (Roth and Murnighan, 1978). At the end of each period, there is a fixed and known probability of 0.97 that the game will continue for another period. This implies that the expected number of periods is approximately 33.[18] Fréchette and Yuksel (2017) demonstrate that the game under random termination, characterized by a continuation rate of r, is theoretically equivalent to an infinitely repeated game for a risk-neutral player, with a discount factor of r.

At the beginning of the experiment, participants are randomly assigned to a group of eight players and play the game with the same partners throughout the experiment. Each participant is assigned a fixed alphabetical player ID (A–H).[19]

4.1.1 CP and RS treatments

In the CP and RS treatments, participants play the public bad network game in each round. In the first stage of each round, they simultaneously and independently decide who to propose links to. In the second stage, after observing the formed network (including bilateral and unilateral link proposals), participants simultaneously choose between actions x and y. Once all decisions are made, each participant's second-stage action is independently detected with a probability of 15%. In CP, only participants detected to have chosen y are punished, while in RS, the detected participants and their neighbors are punished. Punishment reduces the gain component of the affected player's payoff to zero.

In both treatments, participants begin the first stage of round one without any prior

[18] The expected number of periods is calculated by $\frac{1}{1-0.97}$.

[19] As shown in the experimental screenshots in Appendix C, each player is represented by a node in the network graph, with player IDs displayed nearby.

history, only an empty network displaying player IDs. At the end of each round, they receive feedback on that round, including the formed network (with established neighbors and unilateral link proposals), the actions and payoffs of each player in the group, and the detected players. From round two onward, participants can review all feedback from previous rounds.[20] Since player IDs remain fixed throughout the experiment, participants can track others' past behavior, enabling reputation formation.

4.1.2 PI treatment

In the PI treatment, participants repeatedly play the game of the main model with random termination, similar to the RS treatment. However, in contrast to the RS treatment, participants' choices between actions x and y are private information in the PI treatment, and they have the option to choose whether to disclose their actions to all group members.

In each round, following the two decision stages, the computer randomly detects each participant's action with a fixed probability of 15%. After this, participants first receive feedback on that round, which includes the formed network, the participants who are detected for choosing y, the total number of group members who have chosen the public bad action, and their own payoffs.[21]

Following the above feedback, an additional action-disclosure stage is implemented at the end of each round. In this stage, participants decide whether to disclose their actions from the current round at a cost of five units. If disclosed, their actions and payoffs become publicly observable to all group members at the start of the next round; otherwise, they remain hidden and are shown as blank. As in CP and RS treatments, participants in the PI treatment have access to all available information from previous rounds at the beginning of each round. The key difference is that in the PI treatment, the history excludes the actions and payoffs of participants who chose not to disclose them.

[20] Figure 4 in Appendix C provides a screenshot of the decision-making page of the first stage. On the left of this screen, the feedback from the most recent round is displayed as a default and participants can access the feedback from any other previous round.

[21] The disclosure of detected y mirrors real-life practices, where public bad actions are typically made public once detected. For example, regulatory agencies publish reports on restaurants that fail to meet food safety standards, and tax authorities disclose identified cases of tax evasion.

4.2 Procedures

Our experiment was conducted at the Shanghai University of Finance and Economics in 2022. Participants were recruited from the subject pool of the Economic Lab using Ancademy.[22] Each participant was enrolled in only one of the three treatments. We conducted six sessions with four groups in each session. To control for session effects, two treatments were conducted simultaneously in each session. We conducted eight groups of eight participants for each of the three treatments. A total of 192 students participated in the experiment, most of whom were undergraduate students with various majors.

The participants interacted via computer terminals and we programmed the experiment using z-Tree (Fischbacher, 2007). Upon arrival, participants were randomly seated in the laboratory. At the beginning of the experiment, they were required to read the instructions displayed on the computer screen and answer all control questions correctly. At the end of the experiment, participants answered a questionnaire of demographic information. The experimental instructions and screenshots are provided in Appendix C.

To simplify the between-treatment comparisons, we performed random terminations before running the sessions and chose the same set of period numbers across the three treatments. As a result, the randomly realized period number for the eight independent groups for all treatments are 17, 27, 30, 32, 39, 40, 43, and 47.

The participants earned currency points in the experiment, which were calculated as the sum of all points accumulated across all rounds, and the exchange rate is 80 points = 1 Chinese *yuan* (CNY). On average, participants earned 78 CNY (equivalent to about 11 USD), including a show-up fee of 15 CNY (about 2 USD). Each session lasted between 75 and 100 minutes.

4.3 Hypotheses

Our primary goal is to examine the effects of the RS mechanism in our public bad network game, for which we propose a hypothesis regarding the behavioral differences in CP and RS treatments. Then, we propose a hypothesis for the effect of the information environment regarding the behavioral differences in RS and PI treatments.

According to the theoretical analysis in Section 3.2, although noncooperation and full cooperation can be sustained as equilibrium in CP and RS, the basin of attraction theory

[22] Ancademy is a platform for social sciences experiments.

indicates that full cooperation is more likely to arise in RS than in CP. Therefore, we propose the following hypothesis:

Hypothesis 1. *The rate of y choices is lower in the RS treatment than in the CP treatment.*

Next, we develop a hypothesis regarding the effect of the information environment. First, as disclosing information in PI incurs cost, we conjecture that players intending to choose action y (referred to as "defectors") will not reveal their decisions to avoid damaging their reputations. In contrast, players intending to choose action x (referred to as "cooperators") have an incentive to disclose their decisions because it can establish a favorable reputation. Assuming that cooperators consistently choose to disclose their actions, we anticipate that the pattern of public bad action choices in PI will resemble that of RS, as a player's decision not to disclose their choice unequivocally signals a defector type. In summary, the behavior in treatment PI crucially depends on whether cooperators always disclose their actions. The more likely they are to disclose, the smaller the difference of action choices will be in RS and PI.

Hypothesis 2. *In treatment PI, a higher disclosure rate of those choosing x leads to a smaller difference of action choices between PI and RS.*

5 Experimental Results

5.1 Treatment level performances

In the experiment, the number of repetition rounds varies across the eight groups within each treatment, with each group respectively participating in 17, 27, 30, 32, 39, 40, 43, or 47 rounds. To compare participants' behaviors across treatments, we divide the repeated rounds of each group into segments of 10 rounds. Specifically, we analyze participants' behaviors within the following round intervals: 1-10, 11-20, 21-30, 31-40, and 41-50.

Figure 1 presents the evolution of the average cooperation level (represented by the number of players choosing x within each group) and the average number of neighbors across the three treatments. In the top panel, a notable trend is observed in the RS treatment, wherein most groups demonstrate a consistent increase in the average cooperation level, ultimately converging to the highest level of eight after round 30.

Conversely, groups in the CP treatment exhibit fluctuating cooperation levels at a relatively low level without any clear convergence trend. The PI treatment exhibits two distinct evolutionary patterns. In groups 1, 2, 6, and 8 (referred to as PI-C), cooperation levels converge to the maximum level, resembling to the behavior observed in the RS treatment. Conversely, the cooperation levels for the remaining four groups (3, 4, 5, and 7), referred to as PI-NC, fluctuate at a relatively low level, resembling the pattern observed in the CP treatment.

The bottom panel illustrates a positive correlation between the dynamics of neighborhood formation and cooperation levels in RS and PI treatments. In RS and PI-C groups, participants' average numbers of neighbors gradually increase over time, eventually reaching the highest level of seven. In contrast, the PI-NC groups maintain a relatively low number of neighbors throughout the experiment. However, the dynamics of neighborhood formation in the CP treatment does not exhibit a significant correlation with cooperation levels. Notably, participants in the CP treatment quickly converge to the complete networks from the initial rounds of the experiment.

To provide statistical support for these observations, Table 1 presents groups' average cooperation level, the average number of neighbors for each participant, and the results of pairwise Mann-Whitney tests across treatments for rounds 1-10, 11-20, and 21-30, respectively.[23] In the CP treatment, the average cooperation level exhibits a declining trend over time, decreasing from 3.96 in rounds 1-10 to 3.34 in rounds 11-20, and further to 2.87 in rounds 21-30. This decreasing pattern of cooperation level aligns with well-established findings in previous studies on public goods or bads games (Isaac and Walker, 1988a,b; Ledyard, 1995; Chaudhuri, 2011). However, the cooperation levels in the CP treatment do not converge to nearly zero, as has been commonly observed in studies of standard public goods or bads without punishment (van der Heijden and Moxnes, 1999; Fehr and Gächter, 2000; Moxnes and Van der Heijden, 2003; Gurerk et al., 2006; Lugovskyy et al., 2017). This discrepancy could potentially be attributed to two distinctive features of our setting: (i) the presence of centralized punishment, and (ii) the network formation stage requiring mutual consent. We investigate this channel in more detail in Section 5.2.2. In contrast, the average cooperation level in the RS treatment

[23] Table 5 in Appendix D extends Table 1 by presenting the average cooperation level and number of neighbors for rounds 21-40 and 21-50. The averages and significance test results remain largely unchanged compared with rounds 21-30.

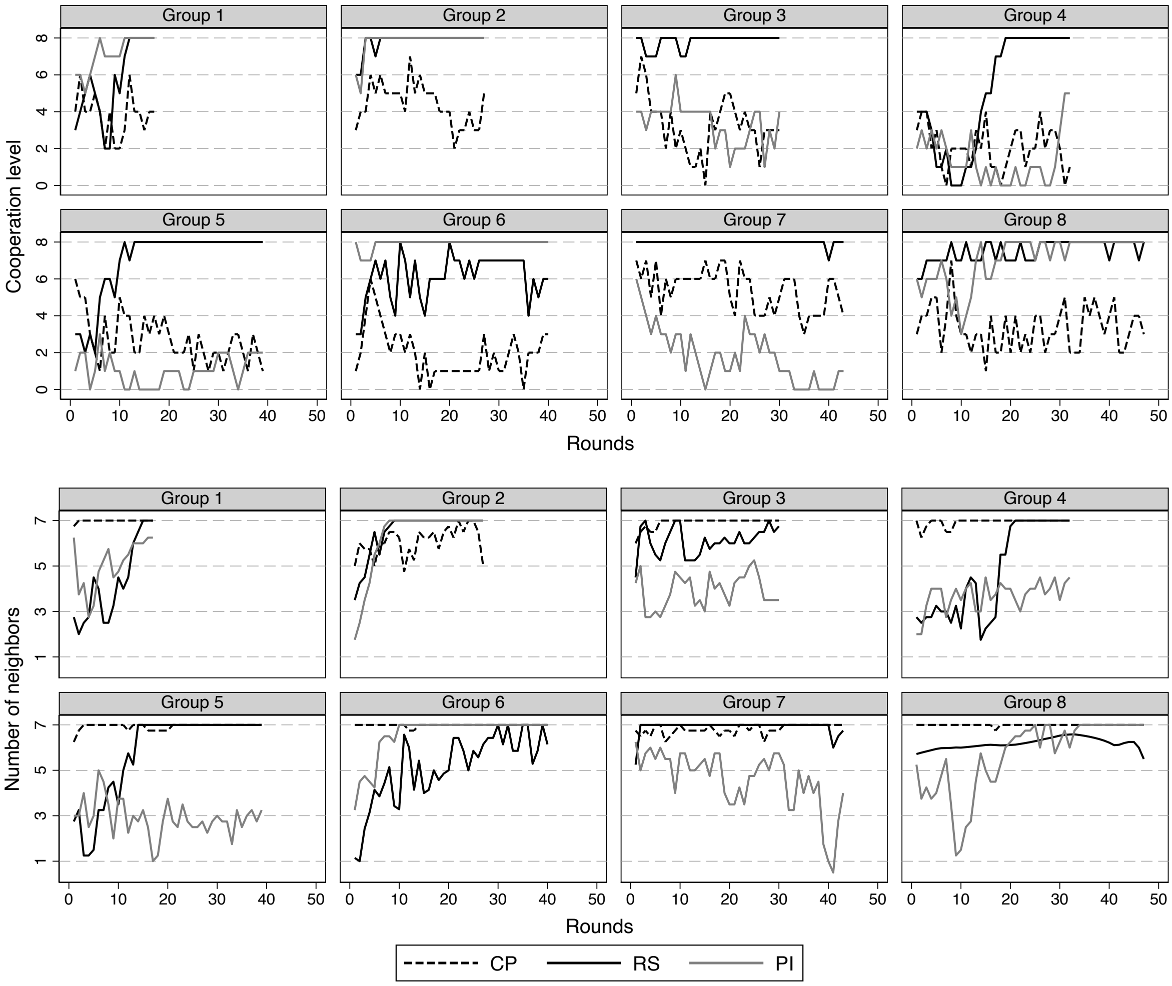


Figure 1: Evolution of groups' cooperation level (top panel) and the number of neighbors (bottom panel).

exhibits an increasing trend over time, rising from 5.71 in rounds 1-10 to 7.23 in rounds 11-20, and further to 7.77 in rounds 21-30. These levels are significantly higher than those in the CP treatment, as indicated by Mann-Whitney tests (rounds 1-10, $p = 0.047$; rounds 11-20, $p = 0.001$; rounds 21-30, $p < 0.001$). In comparison, the PI treatment show no clear trend in either cooperation levels or number of neighbors. The cooperation level in PI is slightly lower than in RS for rounds 21-30 ($p = 0.078$).

The statistics of participants' number of neighbors also support the observations depicted in Figure 1. In the CP treatment, almost all participants quickly form a complete network. In contrast, participants in the RS treatment initially establish fewer neighbors than in CP treatment (rounds 1-10, $p = 0.005$; rounds 11-20, $p = 0.078$). However, over time the number of neighbors steadily rises in RS, reaching nearly seven in rounds 21-30,

Table 1: Statistics of cooperation levels and number of neighbors

	Average cooperation level Maximal level: 8			Average number of neighbors Maximal number: 7		
Rounds	1-10	11-20	21-30	1-10	11-20	21-30
CP	3.96 (1.07)	3.34 (1.64)	2.87 (1.13)	6.75 (0.37)	6.82 (0.33)	6.90 (0.21)
RS	5.71 (2.16)	7.23 (1.19)	7.77 (0.45)	4.57 (1.72)	5.91 (1.03)	6.69 (0.40)
PI	4.79 (2.41)	4.54 (3.47)	4.36 (3.42)	4.34 (0.91)	4.99 (1.57)	5.16 (1.70)
	Mann-Whitney test (p-value)					
CP vs. RS	0.047	0.001	< 0.001	0.005	0.078	0.339
CP vs. PI	0.505	0.591	0.870	< 0.001	0.047	0.054
RS vs. PI	0.367	0.359	0.078	1.000	0.218	0.143

Notes: Standard deviations are in parentheses. Two-sided Mann-Whitney tests are performed at the group level (n = 16 for rounds 1-10 and rounds 11-20; n = 14 for rounds 21-30). Only seven groups remain in rounds 21-30 of each treatment as one group ends at round 17.

which is statistically indistinguishable from CP (p = 0.339). In the PI treatment, the number of neighbors is marginally lower than in RS across all three segments of rounds, although the difference is consistently insignificant.

In summary, these findings demonstrate the substantial impact of the RS mechanism in inhibiting public bad actions when choices are publicly observable, supporting Hypothesis 1. However, when actions remain private, the RS mechanism fosters cooperation in only half of the groups and does not yield significantly higher overall cooperation than CP.

Result 1. *Participants in CP quickly form complete network, but cooperation decreases over time while remaining above zero. Consistent with Hypothesis 1, the RS mechanism significantly reduces public bad actions, increases network size, and improves cooperation levels. In PI, outcomes diverge, with about half the groups converging to full cooperation as in RS, while the remaining groups do not.*

5.2 Determinants of individual-level choices

In this section, we examine the determinants of participants' link decisions and public bad choices across the three treatments. We investigate participants' information disclosure behavior in the PI treatment.

5.2.1 Determinants of link decisions

First, we present the results of regressions using the linear probability model (LPM) with individual and round fixed effects to examine the determinants of link decisions in Table 2.[24] Columns (1)–(3) correspond to each of the three treatments and columns (4)–(5) provide a breakdown of the PI treatment into four successful cooperation groups (PI-C) and the remaining four groups (PI-NC). The dummy dependent variables represent arbitrary participant i's link decisions to arbitrary participant j $(j \in N, j \neq i)$. A value of 1 indicates that participant i proposes the link, while 0 indicates no proposal. The first five explanatory variables represent the joint actions taken by participants i and j in round $t-1$. Specifically, dummy variables $x_{i,t-1}$ and $y_{i,t-1}$ respectively equal 1 if participant i chooses action x and y, and $x_{j,t-1}$, $y_{j,t-1}$, and $nd_{j,t-1}$ equal 1 if participant j is observed choosing action x, y, and does not disclose the action (only in PI), respectively.[25] Note that in columns (3)–(5), if participant j's action can be inferred as choosing action x or y, we let $x_{j,t-1} = 1$ or $y_{j,t-1} = 1$.[26] Next, the explanatory variable, $neighbor_{ij,t-1}$ denotes the link history between participants i and j, equaling 1 if i and j were neighbors in round $t-1$ and 0 otherwise. Finally, the explanatory variable, $reputation_{j,1\sim t-2}$ records player j's rate of observable or inferrable x choices in rounds 1 to $t-2$. In treatments CP and RS, this rate is simply the rate of x choices, whereas in the treatment PI, choice x is observed in the case of disclosing or inferred in the case of group-level full cooperation.

[24] Two fixed-effects dimensions arise in link decisions, the link proposer i and the round t. We employ the LPM approach, which flexibly and efficiently accommodates high-dimensional fixed effects while yielding coefficients with a straightforward and intuitive interpretation.

[25] In columns (1) and (2) of Table 2, the benchmark case (omitted in the regressions) corresponds to the case in which both participants i and j choose action y. In columns (3) to (5), the benchmark case represents the scenario in which participant i chooses action y and participant j does not disclose their action.

[26] Specifically, a few scenarios emerge in which participant j's action can be inferred: (i) when they disclose their actions, (ii) when the feedback information indicates that all others (including participant j) choose x or y, and (iii) participant j is detected to have chosen action y.

Table 2: Determinants of participant i's link decisions in each treatment

	(1) CP	(2) RS	(3) PI	(4) PI-C	(5) PI-NC
$\beta_1 : x_{i,t-1} \times x_{j,t-1}$	0.00604	0.222***	0.226**	0.362***	0.0520
	(0.00335)	(0.0314)	(0.0706)	(0.0235)	(0.0451)
$\beta_2 : x_{i,t-1} \times y_{j,t-1}$	-0.00864	-0.183***	-0.159	-0.320**	-0.113
	(0.00941)	(0.0493)	(0.0904)	(0.0624)	(0.101)
$\beta_3 : y_{i,t-1} \times x_{j,t-1}$	0.0106	0.195***	0.119*	0.336***	0.0137
	(0.00878)	(0.0232)	(0.0585)	(0.0383)	(0.0247)
$\beta_4 : y_{i,t-1} \times y_{j,t-1}$	(benchmark)		-0.0670	-0.156	-0.0580
			(0.0370)	(0.0745)	(0.0366)
$\beta_5 : x_{i,t-1} \times nd_{j,t-1}$			-0.0199	0.0221	-0.00185
			(0.0335)	(0.0600)	(0.0399)
$\beta_6 : y_{i,t-1} \times nd_{j,t-1}$			(benchmark)		
$\beta_7 : neighbor_{ij,t-1}$	0.158**	0.175***	0.262***	0.240***	0.261**
	(0.0471)	(0.0334)	(0.0355)	(0.0281)	(0.0516)
$\beta_8 : reputation_{j,1\sim t-2}$	0.0124	0.213**	0.0977**	0.0788*	0.173*
	(0.00779)	(0.0641)	(0.0333)	(0.0319)	(0.0656)
constant	0.827***	0.422***	0.516***	0.342***	0.580***
	(0.0482)	(0.0636)	(0.0218)	(0.0394)	(0.0269)
N	14504	14504	14504	6888	7616

Standard errors (in parentheses) are clustered at the group level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We first examine whether participants' tendencies to link with others differ based on their action choices in the previous rounds. The significance of variables β_1, β_2, β_3, and β_8 in columns (1) and (2) of Table 2 indicate that participants in the CP treatment who chose action x in the previous round (cooperative participants) and those who chose action y in the previous round (noncooperative participants) exhibit no significant preferences in link decisions toward cooperative or noncooperative others, nor with respect to others' reputations. In other words, participant's link decisions in the CP treatment are independent of others' past behavior. By contrast, all participants in the RS treatment, particularly cooperative ones, propose significantly more links to other cooperative participants than to noncooperative participants and the number of proposals also significantly increases with the reputation levels of others. This indicates that cooperative participants in the RS treatment adopt a *link reciprocity* strategy, making strategic proposal decisions not only based on behavior in the previous round but also on behavior in earlier rounds. This finding is intuitive: due to the responsibility-sharing feature, participants in RS rely on all available information to assess the potential risks of forming links with others. When this concern is absent in CP, participants need not consider others' past behavior since linking with them only brings benefits.

In the PI treatment, when action choices are not disclosed, β_2, β_4, and β_5 are not significantly different from zero, indicating that participants exhibit similar linking tendencies toward noncooperative and nondisclosure participants. Moreover, separately analyzing the PI-C and PI-NC groups reveals that participants in PI-C behave similarly to those in the RS treatment, while participants in PI-NC resemble those in the CP treatment.

These findings indicate that the link reciprocity strategy is rarely adopted in our global public bad games with centralized punishment, in contrast to findings from local public goods/bads games (Cinyabuguma et al., 2005; Page et al., 2005; Rand et al., 2011; Charness and Yang, 2014). This is arguably because linking with others in our CP setting yields only benefits and no costs. By contrast, becoming neighbors with individuals who may choose the public bad action under the RS mechanism can result in potential loss. Therefore, participants must weigh the benefits of any link against its expected costs, making link reciprocity a natural strategy. Furthermore, the distinct linking strategies employed by PI-C and PI-NC participants may help explain overall cooperation disparity between these two groups. We revisit this point after discussing participants' public bad actions and disclosure behavior.

Result 2. *In RS and PI-C, cooperative and noncooperative participants are more likely to link with cooperators than non-cooperators, whereas in CP and PI-NC they show no such tendency. In the PI treatment, participants do not differentiate between non-disclosure and noncooperative participants when forming links.*

5.2.2 Determinants of public bad actions

Next, we analyze the factors influencing participants' decisions to choose public bad action y. Table 3 presents the marginal effects from random-effects probit regressions, examining

Table 3: Determinants of participant i's public bad actions in each treatment

	(1) CP	(2) RS	(3) PI	(4) PI-C	(5) PI-NC
$\beta_1 : y_{i,t-1}$	0.154***	0.091***	0.228***	0.053***	0.266***
	(0.037)	(0.031)	(0.019)	(0.019)	(0.038)
$\beta_2 : \%cooperation_{j\neq i,t-1}$	-0.034***	-0.029***	-0.043***	-0.006***	-0.030**
	(0.010)	(0.006)	(0.006)	(0.002)	(0.013)
$\beta_3 : \#neighbors_{i,t}$	0.100***	-0.015**	-0.030***	-0.019***	-0.033***
	(0.024)	(0.007)	(0.003)	(0.004)	(0.004)
$\beta_4 : \#failed_{i,t}$	0.323***	0.005	0.003	0.000	-0.002
	(0.071)	(0.009)	(0.010)	(0.005)	(0.017)
$\beta_5 : y_{i,t-1}\times\ \#failed_{i,t}$	-0.228***	-0.010	-0.009	-0.007	0.006
	(0.056)	(0.008)	(0.008)	(0.008)	(0.010)
$\beta_6 : detection_{i,t-1}$	0.229***	0.045*	-0.040	0.004	-0.121**
	(0.063)	(0.024)	(0.035)	(0.009)	(0.059)
$\beta_7 : punished^{detection}_{i,t-1}$	-0.086	0.007	-0.037	0.038	-0.027
	(0.067)	(0.025)	(0.035)	(0.033)	(0.068)
$\beta_8 : punished^{neighbor}_{i,t-1}$		-0.009	0.012	0.016	0.004
		(0.020)	(0.028)	(0.011)	(0.031)
N	2136	2136	2136	1016	1120

Standard errors (in parentheses) are clustered at the group level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

participant i's choice of action y in round t for each treatment.[27] The dependent variable is a dummy that equals 1 if participant i chooses action y in round t. The explanatory variables include lagged variables from round $t-1$ such as whether participant i chose action y ($y_{i,t-1}$), other participants' cooperation rate in the group ($\%cooperation_{j\neq i,t-1}$), whether participant i was detected regardless of their choice ($detection_{i,t-1}$), whether they were punished for being detected choosing y ($punished^{detection}_{i,t-1}$), and whether they were punished owing to their neighbors' behavior ($punished^{neighbor}_{i,t-1}$). Additional explanatory variables include the number of neighbors in round t ($\#neighbors_{i,t}$), the number of failed proposals ($\#failed_{i,t}$) in round t, and its interaction with $y_{i,t-1}$.

The regression results demonstrate the significant influence of other participants'

[27] Table 6 in Appendix D presents the results of fixed-effects logit regressions, demonstrating the robustness of Table 3.

cooperation rate (β_2) on an individual's decision to refrain from taking public bad actions across all three treatments. This finding supports the existence of "conditional cooperators," who are more likely to cooperate when they observe others doing so (Keser and Van Winden, 2000; Fischbacher et al., 2001; Kurzban and Houser, 2005; Rustagi et al., 2010; Fischbacher and Gächter, 2010; Yang et al., 2018). Moreover, while cooperative and noncooperative participants in the RS treatment in Table 2 exhibit a significant aversion to establishing links with noncooperators, the number of failed proposals (β_4) does not directly affect participants' choices of public bad actions. Instead, the number of neighbors (β_3) emerges as a significant factor that discourages participants from choosing action y in the RS treatment.

By combining the determinants of participants' link decisions and public bad choices, we conclude that successful cooperation in the RS treatment is attributable to participants' preference for linking with cooperators and avoiding links with noncooperators. This exclusionary behavior deters non-cooperators from choosing the public bad action, particularly when they form more neighbors, as they fear future exclusion and the resulting loss of network benefits. However, when participants do not distinguish between cooperators and noncooperators in link formation, noncooperators no longer face this threat, and refraining from the public bad action is no longer advantageous.

In the CP treatment, apart from the cooperation rate in the previous round (β_2), the only factor that discourages noncooperative participants from choosing action y is their number of failed proposals (β_5). This indicates that noncooperative participants are more likely to switch to cooperation when they experience more failed proposals. Moreover, the larger absolute value of β_4 compared with β_5 indicates that when cooperative participants' proposals fail, they are even more likely to switch to noncooperation. In other words, both cooperative and noncooperative participants adjust their behavior in response to failed proposals, preventing behavioral convergence. The intuition is that, because mutual consent is required to establish neighbors, failed proposals discourage cooperators from continuing to cooperate, while motivating noncooperative participants to start cooperating instead. This switching pattern explains why the the CP treatment' cooperation rate fluctuates over time rather than converging to zero, as is commonly observed in the literature on standard public good games. Furthermore, being punished by the centralized punishment itself does not significantly affect behavior in the CP treatment

(β_7). However, it is possible that participants already consider the probability of being punished when making their decisions, regardless of whether they are actually punished. In summary, we demonstrate that the mutual consent requirement in network formation can explain why behavior fails to converge in the CP treatment, but no direct evidence for the centralized punishment mechanism is found.

In contrast to the RS treatment, the number of neighbors (β_3) in the CP treatment does not discourage participants from choosing the public bad. This finding aligns with the conclusion that neither cooperative nor noncooperative participants in CP show a significant aversion to linking with non-cooperators, as shown in Table 2. Consequently, no mechanism to deter noncooperative participants from choosing action y is evident, which may explain the lower cooperation level observed in CP. Notably, when a participant is detected in round $t-1$ in both CP and RS treatments, it encourages them to choose the public bad in round t (β_6). This implies that being detected *per se* may backfire in our setting. Finally, receiving punishment, regardless of its sources, has no significant effect on cooperation behavior.

In the PI treatment, the determinants of public bad choices do not qualitatively differ from those in the RS treatment, nor between PI-C and PI-NC groups.

Result 3. *Participants exhibit conditional cooperation in all three treatments. In RS and PI treatments, a larger number of neighbors significantly reduces the likelihood of choosing public bad actions, whereas the opposite holds in the CP treatment. The determinants of public bad actions do not differ significantly between PI-C and PI-NC.*

5.2.3 Information disclosure in PI

This section analyzes the dynamics and determinants of information disclosure behavior in the PI treatment and explains PI-C and PI-NC groups' heterogeneous performance. Figure 2 presents each group's disclosure rate evolution alongside cooperation levels for comparison. To enhance clarity, we present PI-C groups in the top panel and PI-NC groups in the bottom panel. The figure reveals a notable disparity in information disclosure between PI-C and PI-NC before convergence. Specifically, in rounds 1-10, participants in PI-C demonstrate a significantly higher disclosure rate of 59% compared with 16% in PI-NC (Mann-Whitney test, $p = 0.029$, $n = 8$). However, once full cooperation is reached, the disclosure rate in PI-C sharply declines to 20% in rounds

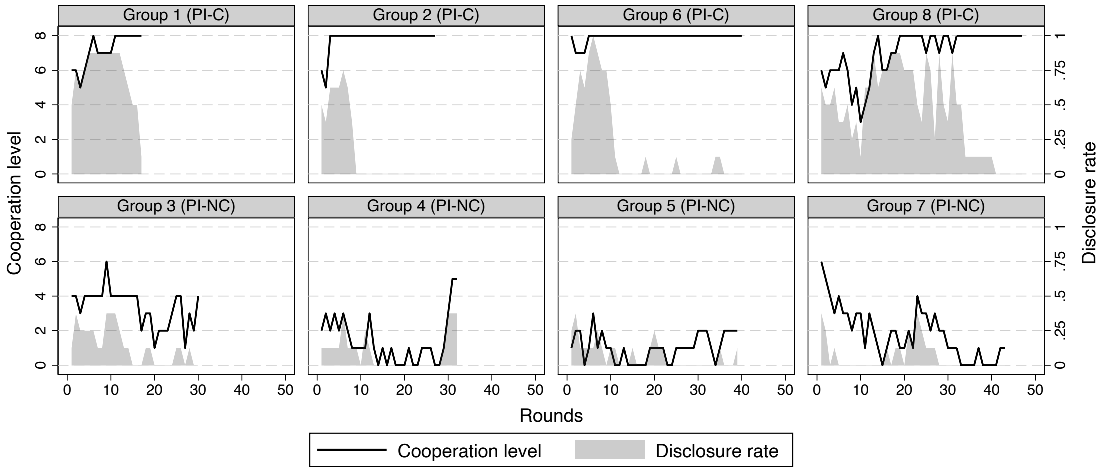


Figure 2: Cooperation level and and disclosure rate evolution in PI

21-30, at which point it no longer differs significantly from PI-NC (Mann-Whitney test, $p = 0.629$, $n = 7$). In rounds where groups have not yet achieved full cooperation, cooperative participants' disclosure rates are significantly higher in PI-C compared with PI-NC (73.0% versus 31.1%; Mann-Whitney test, $p = 0.029$, $n = 8$). Finally, it is noteworthy that 95.4% of all disclosed decisions in the PI treatment were x choices, indicating that the primary purpose of disclosure is reputation formation.

Table 4 presents the marginal effects from random-effects probit regressions that investigate the determinants of participant i's decisions to disclose action x in PI, PI-C, and PI-NC groups.[28] The dependent variable is a binary indicator that equals 1 if participant i discloses action x in round t. The regressions examine the impact of several explanatory variables, encompassing whether participant i disclosed action x in the previous round ($d_{i,t-1}$), the group's disclosure rate of x choices in the previous round ($\%d_{t-1}$), the number of participant i's failed proposals in the current round ($\#failed_{i,t}$). We also introduce a set of dummy variables to capture different cooperation levels in the current round, divided into low ($lowcoop_t$, 0–3 players choose x), high ($highcoop_t$, 4–7 players choose x), and full ($fullcoop_t$, all eight players choose x) cooperation. Further explanatory variables include whether participant i chose action y in the previous round

[28] We specifically analyze disclosure behavior when participants choose to disclose action x, as it represents the majority of disclosed information and is closely correlated with the cooperation success observed in the PI-C group. For a comprehensive examination of the determinants of disclosure behavior irrespective of chosen actions, see Table 7 in Appendix D. In addition, Table 8 in Appendix D presents the results of fixed-effects logit regressions, which confirm the robustness of Table 4.

($y_{i,t-1}$) and participant i's number of neighbors in the current round ($\#neighbors_{i,t}$).

Column (1) of Table 4 shows that individual's disclosure behavior, the rest of the group's disclosure rate in the previous round, the number of failed proposals, and the number of neighbors in the current round all have a significant positive effect on disclosure. The effect of cooperation levels on disclosing exhibits an inverted U-shape, wherein compared with the high cooperation benchmark, the disclosure rate is significantly lower in both low- (β_5) and full-cooperation (β_7) cases.[29] This finding indicates that as cooperation increases, participants initially become more likely to disclose their x choices; however, they reduce disclosure once full cooperation is achieved. Columns (2) and (3) of Table 4 reveal that the positive effects of the past group disclosure rate and the number of neighbors are significant only in PI-NC, while the positive effect of failed proposals is stronger in PI-C than in PI-NC.

Thus far, we have identified two key factors in participants' decision making that distinguish the PI-C and PI-NC groups. First, results in Section 5.2.1 indicate that participants in PI-C exhibit preferences similar to those in the RS treatment when forming connections, whereas those in PI-NC display preferences resembling those in the CP treatment. Second, cooperative participants' disclosure rates are higher in PI-C than in PI-NC. Combining these findings can potentially explain the higher level of cooperation observed in the PI-C group. Cooperative participants' increased disclosure rate makes the decision-making environment in PI-C similar to that of the RS treatment. As a result, cooperative participants are more likely to establish stable and extensive networks with one another, motivating noncooperators to connect with cooperators, refrain from choosing the public bad action, and thereby foster successful cooperation within the PI-C groups.

In summary, our findings reveal the significance of of cooperative participants' early information disclosure in facilitating successful cooperation within the PI treatment. Early disclosure enables cooperators to connect with one another and encourages noncooperators to adopt cooperative behavior. However, once full cooperation is achieved, the role of disclosure is diminished, as participants can sustain cooperation without relying heavily on it.

[29] Because the effect of cooperation levels is non-monotonic, we employ three dummy variables representing different cooperation levels, rather than using a single explanatory variable for cooperation level.

Table 4: Determinants of participant i's disclosure of action x

	(1) PI	(2) PI-C	(3) PI-NC
$\beta_1 : d_{i,t-1}$	0.108^{***}	0.129^{***}	0.108^{***}
	(0.021)	(0.037)	(0.033)
$\beta_2 : \%d_{t-1}$	0.179^{**}	0.163	0.102^{***}
	(0.070)	(0.132)	(0.005)
$\beta_3 : \#failed_{i,t}$	0.028^{***}	0.033^{***}	0.017^{**}
	(0.006)	(0.008)	(0.008)
$\beta_4 : \#neighbors_{i,t}$	0.017^{***}	0.021	0.015^{***}
	(0.005)	(0.017)	(0.005)
$\beta_5 : lowcoop_t$	-0.119^{**}	-0.220^{***}	-0.052^{***}
	(0.047)	(0.027)	(0.007)
$\beta_6 : highcoop_t$		(benchmark)	
$\beta_7 : fullcoop_t$	-0.229^{***}	-0.308^{***}	
	(0.043)	(0.066)	
$\beta_8 : y_{i,t-1}$	-0.032	0.014	-0.013
	(0.020)	(0.071)	(0.013)
N	2136	1016	1120

Standard errors (in parentheses) are clustered at the group level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Result 4. *Cooperative participants' disclosure rate is higher in PI-C than in PI-NC, supporting Hypothesis 2. Information disclosure is essential before cooperation converges to the highest level but becomes unnecessary once full cooperation is achieved.*

5.3 Efficiency

This section analyzes efficiency loss across treatments, defined as the reduction in participants' payoffs relative to the full-cooperation outcome. While the RS mechanism effectively discourages public bad actions, it also introduces additional costs by forming fewer links, particularly in groups and rounds that have not yet reached full cooperation. In addition, responsibility sharing may generate further costs through punishment. Accordingly, we decompose participants' efficiency loss into three components for rounds

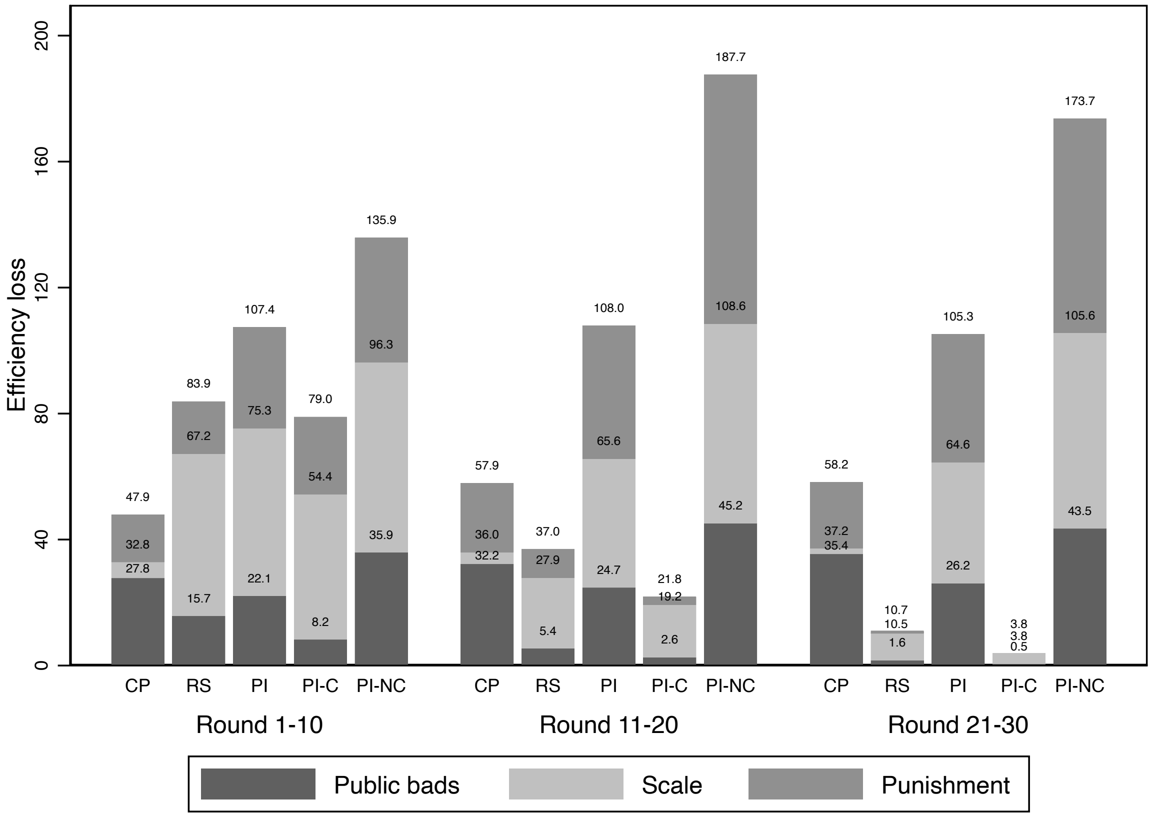


Figure 3: **Decomposition of average efficiency losses for each treatment** (the level of accumulated loss is presented above each bar).

1-10, 11-20, and 21-30, as shown in Figure 3. The first component captures the loss from choosing public bad actions.[30] The second component measures scale loss (i.e., reduction from forming fewer neighbors than the efficient complete network. The third component represents the total punishment imposed on participants.

Efficiency losses in the CP treatment remain steady over time, primarily driven by frequent public bad choices and resulting punishments. In the first 10 rounds of the RS treatment, most efficiency loss comes from the scale component, as participants do not form large networks during these rounds. Notably, the loss from public bad actions is significantly lower in RS than in CP (Mann-Whitney test, $p = 0.047$, $n = 16$). Moreover, although participants share punishment in RS, the efficiency loss from punishments does not differ significantly from that of CP (Mann-Whitney test, $p = 0.878$, $n = 16$). After

[30] This component has two parts. The first is the cost of negative externalities caused by others' public bad actions. The second is the gain from individual's choosing the public bad action. Therefore, we define the loss from public bad actions as the net difference between these two parts.

round 10, efficiency losses from all three components in RS gradually decline, and by rounds 21-30, become significantly lower than in CP (Mann-Whitney test, $p = 0.001$, $n = 14$).

In the PI treatment, overall efficiency loss across the three segments remains high and does not change over time. Dividing the PI treatment into PI-C and PI-NC groups reveals a contrasting pattern. In PI-C, where the RS mechanism successfully promotes cooperation, efficiency loss gradually declines toward zero. In contrast, in PI-NC, where the mechanism fails to improve cooperation, efficiency loss increases over time. Notably, in rounds 21–30, efficiency loss in PI-NC is even significantly greater than in CP (Mann-Whitney test, $p = 0.006$, $n = 11$). A closer comparison between PI-NC and CP reveals no significant difference in the loss from public bad actions (Mann-Whitney test, $p = 0.109$, $n = 11$). Instead, the greater loss in PI-NC arises mainly from two sources, (i) network scale loss due to fewer neighbors (Mann-Whitney test, $p = 0.006$, $n = 11$), potentially because participants in PI-NC shy away from linking with non-cooperators to avoid responsibility sharing; and (ii) greater punishment imposed under the responsibility-sharing rule (Mann-Whitney test, $p = 0.006$, $n = 11$).

Result 5. *Efficiency loss in RS is significantly smaller than in CP. In PI, substantial losses arise when the RS mechanism fails, driven by reduced network size and additional punishments from responsibility sharing.*

6 Conclusion

This study investigates the effect of the responsibility-sharing mechanism on controlling public bad actions employing an infinitely repeated two-stage game in an eight-player economy. In each period, players engage in two stages. First, players endogenously form a network and they decide whether to engage in a public bad action that provides individual benefits but harms other players' welfare. A central institution independently and randomly detects players' action choices. Under the RS mechanism, a player is punished if either they or one of their neighbors is detected to have chosen the public bad action.

Theoretical analysis of the game demonstrates that both noncooperation, where all players choose the public bad action, and full cooperation, where all players refrain from it, yield equilibrium outcomes. However, using the basin of attraction criterion, we determine that full cooperation is more likely to emerge under the RS mechanism. We further demonstrate that these results remain robust across a broad range of parameters.

We also conduct an experiment to test the effect of the RS mechanism on mitigating public bad actions, revealing that the mechanism effectively reduces public bad provisions when participants' actions are publicly known. In the treatment without the mechanism (CP treatment), participants tend to form complete networks and frequently choose the public bad action in each period. In contrast, in the RS treatment, participants tend to avoid linking with non-cooperative players. As a result, noncooperative players refrain from providing the public bad to build links with more players, resulting in full cooperation. In addition, we investigate the effectiveness of the mechanism when players' actions are private information and disclosure is voluntary. Our results indicate that, the mechanism is only partially effective in this setting, with convergence to cooperation depending crucially on the extent of information disclosure.

To the best of our knowledge, this is the first study to investigate the effect and limitations of responsibility sharing among endogenously formed neighbors in the context of public bad. Our research opens several avenues for future inquiry. First, in many real-life scenarios, interactions are formed through centralized institutions rather than decentralized network formation. Further exploration is needed to understand how these institutional differences may affect strategies and choices under RS. Second, while we model RS as equal punishment for all players within a network, other forms of RS —such as alternative punishment allocation rules— can yield different outcomes. Exploring more nuanced design could deepen our understanding of how punishment allocation rules shape cooperation. Finally, regarding information availability, an alternative assumption is that players only observe the actions of their direct neighbors, rather than the entire group as in our experiment. We conjecture that the RS mechanism would still be effective in this case, because players would remain motivated to avoid public bad actions out of concern that their neighbors might sever links. Examining this conjecture, both theoretically and experimentally, is a crucial direction for future research.

References

Acemoglu, Daron and Alexander Wolitzky, "Sustaining cooperation: Community enforcement versus specialized enforcement," *Journal of the European Economic Association*, 2020, *18* (2), 1078–1122.

Ahn, Toh-Kyeong, R Mark Isaac, and Timothy C Salmon, "Coming and going: Experiments on endogenous group sizes for excludable public goods," *Journal of Public Economics*, 2009, *93* (1-2), 336–351.

Aldashev, Gani and Giorgio Zanarone, "Endogenous enforcement institutions," *Journal of Development Economics*, 2017, *128*, 49–64.

Ali, S Nageeb and David A Miller, "Enforcing cooperation in networked societies," *working paper*, 2014.

Anderson, Lisa R and Sarah L Stafford, "Punishment in a regulatory setting: Experimental evidence from the VCM," *Journal of Regulatory Economics*, 2003, *24* (1), 91–110.

Andreoni, James, "Why free ride?: Strategies and learning in public goods experiments," *Journal of Public Economics*, 1988, *37* (3), 291–304.

_ **and Laura K Gee**, "Gun for hire: Delegated enforcement and peer punishment in public goods provision," *Journal of Public Economics*, 2012, *96* (11-12), 1036–1046.

Attanasio, Orazio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons, and Heike Harmgart, "Group lending or individual lending? Evidence from a randomised field experiment in Mongolia," Technical Report, WZB Discussion Paper 2014.

Bala, Venkatesh and Sanjeev Goyal, "A noncooperative model of network formation," *Econometrica*, 2000, *68* (5), 1181–1229.

Bó, Pedro Dal and Guillaume R Fréchette, "The evolution of cooperation in infinitely repeated games: Experimental evidence," *American Economic Review*, 2011, *101* (1), 411–429.

Bó, Pedro Dal, Andrew Foster, and Louis Putterman, "Institutions and behavior: Experimental evidence on the effects of democracy," *American Economic Review*, 2010, *100* (5), 2205–2229.

Boosey, Luke and R Mark Isaac, "Asymmetric network monitoring and punishment in public goods experiments," *Journal of Economic Behavior & Organization*, 2016, *132*, 26–41.

Brekke, Kjell Arne, Karen Evelyn Hauge, Jo Thori Lind, and Karine Nyborg, "Playing with the good guys. A public good game with endogenous group formation," *Journal of Public Economics*, 2011, *95* (9-10), 1111–1118.

Carpenter, Jeffrey, Shachar Kariv, and Andrew Schotter, "Network architecture, cooperation and punishment in public good experiments," *Review of Economic Design*, 2012, *16* (2), 93–118.

Cason, Timothy N and Lata Gangadharan, "Promoting cooperation in nonlinear social dilemmas through peer punishment," *Experimental Economics*, 2015, *18*, 66–88.

Charness, Gary and Chun-Lei Yang, "Starting small toward voluntary formation of efficient large groups in public goods provision," *Journal of Economic Behavior & Organization*, 2014, *102*, 119–132.

Chaudhuri, Ananish, "Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature," *Experimental Economics*, 2011, *14* (1), 47–83.

Chen, Yan, "Incentive-compatible mechanisms for pure public goods: A survey of experimental research," *Handbook of Experimental Economics Results*, 2008, *1*, 625–643.

Cinyabuguma, Matthias, Talbot Page, and Louis Putterman, "Cooperation under the threat of expulsion in a public goods experiment," *Journal of Public Economics*, 2005, *89* (8), 1421–1435.

Cornes, Richard and Todd Sandler, *The theory of externalities, public goods, and club goods*, Cambridge University Press, 1996.

DeAngelo, Gregory and Laura K Gee, "Peers or police?: The effect of choice and type of monitoring in the provision of public goods," *Games and Economic Behavior*, 2020, *123*, 210–227.

Deb, Joyee, "Cooperation and community responsibility," *Journal of Political Economy*, 2020, *128* (5), 1976–2009.

Dixit, Avinash, "Trade expansion and contract enforcement," *Journal of Political Economy*, 2003, *111* (6), 1293–1317.

Ebel, Holger and Stefan Bornholdt, "Coevolutionary games on networks," *Physical Review E*, 2002, *66* (5), 056118.

Ellison, Glenn, "Cooperation in the prisoner's dilemma with anonymous random matching," *The Review of Economic Studies*, 1994, *61* (3), 567–588.

Falk, Armin and Michael Kosfeld, "It's all about connections: Evidence on network formation," *Review of Network Economics*, 2012, *11* (3).

Fan, Yiran, "The interaction of bankers' asset and liability management with liquidity concerns," *Journal of Political Economy*, 2021, *129* (8), 2233–2274.

Fehr, Ernst and Simon Gächter, "Cooperation and punishment in public goods experiments," *American Economic Review*, 2000, *90* (4), 980–994.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 2007, *10* (2), 171–178.

_ and Simon Gächter, "Social preferences, beliefs, and the dynamics of free riding in public goods experiments," *American Economic Review*, 2010, *100* (1), 541–556.

_ , _ , and Ernst Fehr, "Are people conditionally cooperative? Evidence from a public goods experiment," *Economics Letters*, 2001, *71* (3), 397–404.

Fishman, Michael J, Jonathan A Parker, and Ludwig Straub, "A dynamic theory of lending standards," Technical Report, National Bureau of Economic Research 2020.

Fowler, James H and Nicholas A Christakis, "Cooperative behavior cascades in human social networks," *Proceedings of the National Academy of Sciences*, 2010, *107* (12), 5334–5338.

Fréchette, Guillaume R and Sevgi Yuksel, "Infinitely repeated games in the laboratory: Four perspectives on discounting and random termination," *Experimental Economics*, 2017, *20* (2), 279–308.

Gächter, Simon and Benedikt Herrmann, "Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2009, *364* (1518), 791–806.

_ and Christian Thöni, "Social learning and voluntary cooperation among like-minded people," *Journal of the European Economic Association*, 2005, *3* (2-3), 303–314.

Gallo, Edoardo and Chang Yan, "The effects of reputational and social knowledge on cooperation," *Proceedings of the National Academy of Sciences*, 2015, *112* (12), 3647–3652.

Goeree, Jacob K, Arno Riedl, and Aljaž Ule, "In search of stars: Network formation among heterogeneous agents," *Games and Economic Behavior*, 2009, *67* (2), 445–466.

Goyal, Sanjeev, Penélope Hernández, Guillem Martínez-Cánovas, Frédéric Moisan, Manuel Muñoz-Herrera, and Angel Sánchez, "Integration and diversity," *Experimental Economics*, 2021, *24* (2), 387–413.

Gracia-Lázaro, Carlos, Alfredo Ferrer, Gonzalo Ruiz, Alfonso Tarancón, José A Cuesta, Angel Sánchez, and Yamir Moreno, "Heterogeneous networks do not promote cooperation when humans play a Prisoner's Dilemma," *Proceedings of the National Academy of Sciences*, 2012, *109* (32), 12922–12926.

Greif, Avner, "Reputation and coalitions in medieval trade: Evidence on the Maghribi traders," *The Journal of Economic History*, 1989, *49* (4), 857–882.

_ , "Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition," *The American Economic Review*, 1993, pp. 525–548.

_ , Paul Milgrom, and Barry R Weingast, "Coordination, commitment, and enforcement: The case of the merchant guild," *Journal of Political Economy*, 1994, *102* (4), 745–776.

Gurerk, Ozgur, Bernd Irlenbusch, and Bettina Rockenbach, "The competitive advantage of sanctioning institutions," *Science*, 2006, *312* (5770), 108–111.

Hanaki, Nobuyuki, Alexander Peterhansl, Peter S Dodds, and Duncan J Watts, "Cooperation in evolving social networks," *Management Science*, 2007, *53* (7), 1036–1050.

He, Simin and Xinlu Zou, "Public goods provision in a network formation game," *Journal of Economic Behavior & Organization*, 2024, *218*, 104–131.

Herrmann, Benedikt, Christian Thoni, and Simon Gachter, "Antisocial punishment across societies," *Science*, 2008, *319* (5868), 1362–1367.

Hiller, Timo, "Peer effects in endogenous networks," *Games and Economic Behavior*, 2017, *105*, 349–367.

Isaac, R Mark and James M Walker, "Communication and free-riding behavior: The voluntary contribution mechanism," *Economic Inquiry*, 1988, *26* (4), 585–608.

_ **and** _ , "Group size effects in public goods provision: The voluntary contributions mechanism," *The Quarterly Journal of Economics*, 1988, *103* (1), 179–199.

Jackson, Matthew O and Alison Watts, "The evolution of social and economic networks," *Journal of Economic Theory*, 2002, *106* (2), 265–295.

_ **and Asher Wolinsky**, "A strategic model of social and economic networks," *Journal of Economic Theory*, 1996, *71* (1), 44–74.

_ **and Yves Zenou**, "Games on networks," in "Handbook of Game Theory with

Economic Applications," Vol. 4, Elsevier, 2015, pp. 95–163.

Kamijo, Yoshio, Tsuyoshi Nihonsugi, Ai Takeuchi, and Yukihiko Funaki, "Sustaining cooperation in social dilemmas: Comparison of centralized punishment institutions," *Games and Economic Behavior*, 2014, *84*, 180–195.

Kandori, Michihiro, "Social norms and community enforcement," *The Review of Economic Studies*, 1992, *59* (1), 63–80.

Keser, Claudia and Frans Van Winden, "Conditional cooperation and voluntary contributions to public goods," *Scandinavian Journal of Economics*, 2000, *102* (1), 23–39.

Kurzban, Robert and Daniel Houser, "Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations," *Proceedings of the National Academy of Sciences*, 2005, *102* (5), 1803–1807.

Ledyard, John O, "Public goods: A survey of experimental research," *The Handbook of Experimental Economics*, 1995, *(Chap.2).*

Leibbrandt, Andreas, Abhijit Ramalingam, Lauri Sääksvuori, and James M Walker, "Incomplete punishment networks in public goods games: Experimental evidence," *Experimental Economics*, 2015, *18* (1), 15–37.

Li, Xuelong, Marko Jusup, Zhen Wang, Huijia Li, Lei Shi, Boris Podobnik, H Eugene Stanley, Shlomo Havlin, and Stefano Boccaletti, "Punishment diminishes the benefits of network reciprocity in social dilemma experiments," *Proceedings of the National Academy of Sciences*, 2018, *115* (1), 30–35.

Lugovskyy, Volodymyr, Daniela Puzzello, Andrea Sorensen, James Walker, and Arlington Williams, "An experimental study of finitely and infinitely repeated linear public goods games," *Games and Economic Behavior*, 2017, *102*, 286–302.

Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval, "Monetary and nonmonetary punishment in the voluntary contributions mechanism," *American Economic Review*, 2003, *93* (1), 366–380.

Masten, Scott E and Jens Prüfer, "On the evolution of collective enforcement institutions: Communities and courts," *The Journal of Legal Studies*, 2014, *43* (2), 359–400.

Morduch, Jonathan, "The microfinance promise," *Journal of Economic Literature*, 1999, *37* (4), 1569–1614.

Moxnes, Erling and Eline Van der Heijden, "The effect of leadership in a public bad experiment," *Journal of Conflict Resolution*, 2003, *47* (6), 773–795.

Nikiforakis, Nikos, "Punishment and counter-punishment in public good games: Can we really govern ourselves?," *Journal of Public Economics*, 2008, *92* (1-2), 91–112.

_ and Hans-Theo Normann, "A comparative statics analysis of punishment in public-good experiments," *Experimental Economics*, 2008, *11*, 358–369.

Oei, Shu-Yi, "The offshore tax enforcement dragnet," *EmOry lJ*, 2017, *67*, 655.

_ and Diane Ring, "Leak-driven law," *UCLA L. Rev.*, 2018, *65*, 532.

Ostrom, Elinor, *Governing the commons: The evolution of institutions for collective action*, Cambridge University Press, 1990.

_ , James Walker, and Roy Gardner, "Covenants with and without a sword: Self-governance is possible," *American Political Science Review*, 1992, *86* (2), 404–417.

Page, Talbot, Louis Putterman, and Bulent Unel, "Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency," *The Economic Journal*, 2005, *115* (506), 1032–1053.

Putterman, Louis, Jean-Robert Tyran, and Kenju Kamei, "Public goods and voting on formal sanction schemes," *Journal of Public Economics*, 2011, *95* (9-10), 1213–1222.

Rand, David G, Samuel Arbesman, and Nicholas A Christakis, "Dynamic social networks promote cooperation in experiments with humans," *Proceedings of the National Academy of Sciences*, 2011, *108* (48), 19193–19198.

Riedl, Arno, Ingrid MT Rohde, and Martin Strobel, "Efficient coordination in weakest-link games," *The Review of Economic Studies*, 2016, *83* (2), 737–767.

_ , _ , and _ , "Free neighborhood choice boosts socially optimal outcomes in stag-hunt coordination problem," *Scientific Reports*, 2021, *11* (1), 1–12.

Roth, Alvin E and J Keith Murnighan, "Equilibrium behavior and repeated play of the prisoner's dilemma," *Journal of Mathematical Psychology*, 1978, *17* (2), 189–198.

Rustagi, Devesh, Stefanie Engel, and Michael Kosfeld, "Conditional cooperation and costly monitoring explain success in forest commons management," *Science*, 2010, *330* (6006), 961–965.

Santos, Francisco C, Jorge M Pacheco, and Tom Lenaerts, "Cooperation prevails when individuals adjust their social ties," *PLoS Computational Biology*, 2006, *2* (10), e140.

Shirado, Hirokazu, Feng Fu, James H Fowler, and Nicholas A Christakis, "Quality versus quantity of social ties in experimental cooperative networks," *Nature Communications*, 2013, *4* (1), 2814.

Shitovitz, Benyamin and Menahem Spiegel, "Cournot-Nash and Lindahl equilibria in pure public "bad" economies," *Economic Theory*, 2003, *22* (1), 17–31.

Song, Yangbo and Mihaela van der Schaar, "Dynamic network formation with incomplete information," *Economic Theory*, 2015, *59* (2), 301–331.

Takahashi, Satoru, "Community enforcement when players observe partners' past play," *Journal of Economic Theory*, 2010, *145* (1), 42–62.

Teteryatnikova, Mariya and James Tremewan, "Myopic and farsighted stability in network formation games: An experimental study," *Economic Theory*, 2020, *69* (4), 987–1021.

The Pie News, "UK universities withdraw offers after Pearson cheating concerns," `https://thepienews.com/news/uk-universities-pearson-cheating/` 2023.

Tyran, Jean-Robert and Lars P Feld, "Achieving compliance when legal sanctions are non-deterrent," *The Scandinavian Journal of Economics*, 2006, *108* (1), 135–156.

van der Heijden, Eline CM and Erling Moxnes, "Information feedback in public-bad games: A cross-country experiment," *Tilburg University, Center for Economic Research, Discussion Paper*, 1999.

van Leeuwen, Boris, Abhijit Ramalingam, David Rojo Arjona, and Arthur Schram, "Centrality and cooperation in networks," *Experimental Economics*, 2019, *22*, 178–196.

Wolitzky, Alexander, "Cooperation with network monitoring," *Review of Economic Studies*, 2013, *80* (1), 395–427.

Xiao, Erte and Daniel Houser, "Punish in public," *Journal of Public Economics*, 2011, *95* (7-8), 1006–1017.

Yang, Chun-Lei, Boyu Zhang, Gary Charness, Cong Li, and Jaimie W Lien, "Endogenous rewards promote cooperation," *Proceedings of the National Academy of Sciences*, 2018, *115* (40), 9968–9973.

Appendix A The Generalized Model

This section introduces the generalized model and provides the corresponding equilibrium analysis. In the generalized model, we relax the assumption of costless link cost between neighbors in Section 3, and allow for the cost to be zero or positive.

A. 1 Model

A. 1.1 Public bad network game

Let $N = \{1, 2, .., n\}$ be a set of players. In the first stage, players simultaneously make link proposals $\mathbf{g}_i = \{g_{i1}, g_{i2}, ..., g_{i8}\}$, where $g_{ij} = 1$ if player i proposes a link to player j and $g_{ij} = 0$ otherwise, and form a directed network $\mathbf{g} = \{\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_n\}$, with two players becoming neighbors if they mutually propose to one another. Denote $N_i(\mathbf{g}) = \{j \in N | g_{ij}g_{ji} = 1, j \neq i\}$ as the set of player i's neighbors, and $|N_i(\mathbf{g})|$ as the number of i's neighbors. Similar to the assumption in Section 3, link proposals are still costless ($c = 0$); however, for each neighbor $g_{ij}g_{ji} = 1$, both player i and j incur a neighbor link cost of l ($l \geq 0$). This assumption is broadly adopted in the literature on network formation with mutual consent (Jackson and Wolinsky, 1996; Song and van der Schaar, 2015; Hiller, 2017). In addition to the cost of forming neighbors, there is a scale effect associated with having more neighbors, where each neighbor yields a benefit of $\beta > 0$. Hence, player i receives an additional benefit of $\beta|N_i(\mathbf{g})|$ for having $|N_i(\mathbf{g})|$ neighbors.

In the second stage, players simultaneously choose an action $a_i \in \{x, y\}$. Let $\mathbf{a} = (a_1, a_2, ..., a_n)$ denote all players' action profiles. The benefit from choosing action x is u_x, and the benefit from choosing action y is u_y, with $u_x < u_y$. Besides, choosing action y generates negative externalities, imposing a cost of b ($b > 0$) on each other player in the group, while choosing action x does not affect the payoffs of others. We assume that $u_x > u_y - (n-1)b$, ensuring that choosing x is socially efficient.

Therefore, player i's final payoff from the two-stage game, $U_i(\mathbf{g}, \mathbf{a})$, is represented as follows:

$$U_i(\mathbf{g}, \mathbf{a}) = U_{i,gain} - U_{i,loss} - l \sum_{j \in N, k \neq i} g_{ij}g_{ji}, \tag{6}$$

where $U_{i,gain}$ is the gain component, which is the sum of the benefits from the network scale and the action in the second stage, as follows:

$$U_{i,gain} = \beta|N_i(\mathbf{g})| + u_x(u_y), \text{if } a_i = x(y)\ , \tag{7}$$

and $U_{i,loss}$ is the public bad loss imposed by the other players who choose y in the group, as follows:

$$U_{i,loss} = b \sum_{j \in N, k \neq i} \mathbb{I}\{a_j = y\}. \tag{8}$$

A. 1.2 Baseline model: CP with random detection

In the baseline model, the central institution randomly detects each player's actions with a fixed detection probability. If a player chooses y and is detected, they incur a fixed punishment, setting their gain component of the payoff to u_p, where $u_p < u_x$. No punishment is imposed if a player is not detected choosing y. The gain component of player i's payoff in the baseline model is represented as follows:

$$U_{i,gain} = \begin{cases} \beta|N_i(\mathbf{g})| + u_x(u_y), & \text{if } i \text{ is not detected to choose } y \text{ and } a_i = x(y) \\ u_p, & \text{if } i \text{ is detected to choose } y \end{cases}. \tag{9}$$

Detection is exogenous and independent across players, with each player being detected with a fixed probability of α. Additionally, we assume $\alpha \in (0, \frac{u_y - u_x}{u_y - u_p - (n-1)\beta})$, ensuring that the detection rate is sufficiently low such that choosing x is not a dominant strategy. Note that the detection and punishment mechanism only affect the gain component of the payoff function.

A. 1.3 Main model: RS mechanism

In the main model, under the RS mechanism, if player i is detected to choose y, then player i is punished with a lower gain and all of their neighbors also incur the same punishment. Similarly, the RS mechanism punishment only affects the gain component of the payoff function, which is represented as follows:

$$U_{i,gain} = \begin{cases} \beta|N_i(\mathbf{g})| + u_x(u_y), & \text{if neither } i \text{ nor any player in } N_i(\mathbf{g}) \text{ is detected} \\ & \text{to choose } y, \text{ and } a_i = x(y) \\ u_p, & \text{if } i \text{ or at least one player in } N_i(\mathbf{g}) \text{ is detected} \\ & \text{to choose } y \end{cases}. \tag{10}$$

A. 2 Equilibrium analysis

A. 2.1 Equilibria of the one-shot game

The equilibrium of the one-shot game in the generalized model exhibits similar properties to those discussed in Section 3.2. In baseline and main models, choosing action y remains the strictly dominant strategy in the second stage, resulting in a unique equilibrium of the subgame where all players choose y.

Proposition A.1 presents the equilibrium properties of the baseline model. Under the assumption of a sufficiently low neighbor link cost, any network structure that does not involve unilateral proposals can exist in equilibrium in the baseline model.

Proposition A.1. *Suppose $l \leq \beta(1-\alpha)$. For every subgame-perfect equilibrium of the baseline model, $(\boldsymbol{g}, \boldsymbol{a})$, the following properties hold for all $i, j \in N$ with $i \neq j$: (i) $g_{ij} = g_{ji}$ and (ii) $a_i = y$.*

Proof of Proposition A.1. We first prove the strict dominance of action y given any formed network $\mathbf{g}$. Let M denote the number of other players (excluding player i) who choose the action y. Given the assumption of a low detection rate such that $\alpha < \frac{u_y - u_x}{u_y - u_p - (n-1)\beta}$, the expected payoff for any player $i \in N$ when choosing action y is strictly greater than the expected payoff of choosing action x for any network $\mathbf{g}$ and any value of M, regardless of the actions of other players:

$$(1-\alpha)(u_y + \beta|N_i(\mathbf{g})|) + \alpha u_p - bM - l \sum_{j \in N, k \neq i} g_{ij} g_{ji} > u_x + \beta|N_i(\mathbf{g})| - bM - l \sum_{j \in N, k \neq i} g_{ij} g_{ji}.$$

Next, we demonstrate that any network without unilateral proposals, $\mathbf{g}$, can be supported in equilibrium. Under the assumption that $l \leq \beta(1-\alpha)$, no player $i \in N$ has an incentive to break down the links with established neighbors. This is evident from the following inequality, which holds for all $1 \leq k \leq |N_i(\mathbf{g})|$, where k is the number of severed links:

$$\begin{aligned}(1-\alpha)\left(u_y + \beta|N_i(\mathbf{g})|\right) + \alpha u_p - bM - l|N_i(\mathbf{g})| \geq (1-\alpha)\left[u_y + \beta\left(|N_i(\mathbf{g})| - k\right)\right] \\ + \alpha u_p - bM - l\left(|N_i(\mathbf{g})| - k\right).\end{aligned} \tag{11}$$

Furthermore, any network with unilateral proposals cannot be supported in equilibrium. If a unilateral proposal to player i exists in equilibrium, player i will always have an incentive to pay for an additional neighbor, as demonstrated by Equation (11) as well. □

Next, Proposition A.2 outlines the equilibrium properties of the main model. Assuming a sufficiently low neighbor link cost, the equilibrium network is characterized by a maximum number of neighbors for each player, denoted $\hat{m}$, which maximizes the expected payoff when all players choose action y. Additionally, the second condition in Proposition A.2 asserts that no player can propose unilateral links to others with fewer than $\hat{m}$ neighbors, while unilateral proposals to players with $\hat{m}$ neighbors can exist in equilibrium.

Proposition A.2. *Let* $\hat{m} = \arg\max_{m \in N}\{(1-\alpha)^{m+1}(u_y + m\beta) + [1-(1-\alpha)^{m+1}]u_p - lm\}$, *and suppose* $l \leq (1-\alpha)^{\hat{m}}\{ln(1-\alpha)[u_y + (\hat{m}-1)\beta - u_p] + \beta\}$. *For every subgame-perfect equilibrium of the main model,* $(\boldsymbol{g}, \boldsymbol{a})$, *the following properties hold for all* $i, j \in N$ *with* $i \neq j$*: (i)* $|N_i(\boldsymbol{g})| \leq \hat{m}$, *(ii) if* $|N_i(\boldsymbol{g})| < \hat{m}$, $g_{ji} = 0$, *and (iii)* $a_i = y$.

Proof of Proposition A.2. We begin by proving the strict dominance of action y, given any formed network $\mathbf{g}$. Let M denote the number of other players (excluding player i) who choose action y, and let m be the number of player i's neighbors who choose action y. Under the assumption that $\alpha < \frac{u_y - u_x}{u_y - u_p - (n-1)\beta}$, the expected payoff for player i from choosing y is always greater than that from choosing action x, for any network $\mathbf{g}$ and for any values of m and M, as demonstrated by the following inequality:

$$(1-\alpha)^{m+1}\left(u_y + \beta|N_i(\mathbf{g})|\right) + \left[1-(1-\alpha)^{m+1}\right]u_p - bM - l\sum_{j\in N, j\neq i} g_{ij}g_{ji}$$
$$> (1-\alpha)^{m}\left(u_x + \beta|N_i(\mathbf{g})|\right) + \left[1-(1-\alpha)^{m}\right]u_p - bM - l\sum_{j\in N, j\neq i} g_{ij}g_{ji}.$$

Thus, in equilibrium, all players choose y in the second stage.

Next, by the definition of $\hat{m}$, it is evident that no player has an incentive to form more neighbors than $\hat{m}$. Otherwise, they could increase their expected payoff by unilaterally severing some of their existing neighbors. We now demonstrate that forming less than $\hat{m}$ neighbors can also be supported in equilibrium. We first restrict our attention to networks without unilateral proposals. Under the assumption of a relatively low link cost such that, $l < (1-\alpha)^{\hat{m}}\{ln(1-\alpha)[u_y + (\hat{m}-1)\beta - u_p] + \beta\}$, the expected payoff function for forming m neighbors,

$$(1-\alpha)^{m+1}(u_y + m\beta) + [1-(1-\alpha)^{m+1}]u_p - lm,$$

is increasing in m when $m < \hat{m}$ (with the loss from others' public actions ignored, as it does not vary with m). In this case, no player has an incentive to sever links when they have formed m neighbors, as doing so would decrease their expected payoff. Therefore, any network $\mathbf{g}$ without unilateral proposals, where $|N_i(\mathbf{g})| < \hat{m}$ for any player i, can be supported in equilibrium.

Having established that any network without unilateral proposals, where each player forms at most $\hat{m}$ neighbors, can be supported in equilibrium, we finally consider networks with unilateral proposals. It is clear that no unilateral proposal to players with fewer than $\hat{m}$ neighbors can exist in equilibrium, as such players have an incentive to form additional neighbors to increase their expected payoff. However, unilateral proposals to players with exactly $\hat{m}$ neighbors can exist in equilibrium, as these players have no incentive to form additional neighbors. □

Note that both Proposition A.1 and Proposition A.2 impose restrictions on the link cost l. Since the inequality $(1-\alpha)^{\hat{m}}\{ln(1-\alpha)[u_y + (\hat{m}-1)\beta - u_p] + \beta\} \leq \beta(1-\alpha)$ holds for all $\hat{m} \geq 1$, the link cost l must be set sufficiently low, such that

$$l \leq (1-\alpha)^{\hat{m}}\{ln(1-\alpha)[u_y + (\hat{m}-1)\beta - u_p] + \beta\},$$

to satisfy the conditions in Proposition A.1 and Proposition A.2.

A. 2.2 Eliminating weakly dominated strategies

In this section, we apply the elimination of weakly dominated strategies to refine one-shot equilibria in baseline and main models. Proposition A.3 shows that, given a sufficiently low link cost, the unique undominated equilibrium in the baseline model is characterized by players forming a complete network and choosing action y. Conversely, in the main model, the equilibrium involves each player forming $\hat{m}$ neighbors and choosing action y.

Proposition A.3. *Suppose $l \leq (1-\alpha)^{\hat{m}}\{ln(1-\alpha)[u_y + (\hat{m}-1)\beta - u_p] + \beta\}$. After eliminating weakly dominated strategies, the baseline model has a unique, undominated subgame-perfect equilibrium, $(\boldsymbol{g}, \boldsymbol{a})$, where for all $i, j \in N$ with $i \neq j$: (i) $g_{ij} = 1$, and (ii) $a_i = y$, and the undominated subgame-perfect equilibrium in the main model, $(\boldsymbol{g}, \boldsymbol{a})$, satisfies the following properties for all $i \in N$: (i) $|N_i(\boldsymbol{g})| = \hat{m}$, and (ii) $a_i = y$.*

Proof of Proposition A.3. We first consider the baseline model and prove the strategy in which player i proposes $k < n-1$ links and chooses action y is weakly dominated by the strategy in which player i proposes $n-1$ neighbors and chooses action y. Given any profile of other players' strategies, let m denote the number of neighbors that player i establishes when proposing k links, and let $\bar{m}$ denote the number of neighbors for player i when proposing $N-1$ links. Notably, the condition $\bar{m} \geq m$ always holds because $k < n-1$. Then we have the following inequality, which holds for all $m < n-1$, regardless of other players' actions:

$$(1-\alpha)(u_y + \beta\bar{m}) + \alpha u_p - bM - l\bar{m} \geq (1-\alpha)(u_y + \beta m) + \alpha u_p - bM - lm,$$

where M denotes the number of other players (excluding player i) who choose action y.

Next, we prove that, in the main model, the strategy in which player i proposes $k < \hat{m}$ links and chooses action y is weakly dominated by the strategy in which player i proposes $\hat{m}$ links and chooses action y. We have already shown that, the function of the expected payoff (with the loss component omitted),

$$(1-\alpha)^{m+1}(u_y + m\beta) + [1 - (1-\alpha)^{m+1}]u_p - lm,$$

is increasing in the number of neighbors m when $m \leq \hat{m}$. Since proposing $k < \hat{m}$ links results in a lower or equal number of established neighbors compared to proposing $\hat{m}$ links, the expected payoff in the former case is lower or equal to that in the latter case. As a result, the strategy of proposing $k < \hat{m}$ links is weakly dominated by the strategy of proposing $\hat{m}$ links. □

A. 2.3 Equilibria of the infinitely repeated game

In this section, we derive the equilibrium for the infinitely repeated public bad network game with a discount factor δ. First, as illustrated in Proposition 5, the noncooperation strategy, wherein players play the equilibrium of the one-shot game in each period, constitutes a subgame-perfect equilibrium in the infinitely repeated game.

Subsequently, we consider the grim trigger strategy as defined in Definition 1, with one modification: In item (iii) of Definition 1, each player proposes $\hat{m}$ links, thereby forming a fixed $\hat{m}$-regular network, instead of a 1-regular network, after any deviation. We prove that this modified grim trigger strategy constitutes an equilibrium of the infinitely repeated game in both models. As a result, the full-cooperation outcome, where all players form a complete network and choose action x, can still be sustained in the infinitely repeated game. Specifically, in the baseline model, the cost of forming neighbors, l, does not impact the condition on δ for equilibrium, as players consistently form a complete network under the grim trigger strategy. However, in the main model, an increase in the cost l results in a stricter condition on δ, since players form fewer neighbors in response to deviations from the full-cooperation outcome, thereby reducing the cost associated with neighbor formation.

Proposition A.4. *Consider the infinite repetition of the game with a discount factor δ. When $\delta \geq \frac{(1-\alpha)u_y - u_x + \alpha u_p - \alpha(n-1)\beta}{(n-1)b}$, the grim trigger strategy constitutes a subgame perfect equilibrium of the infinite repeated game of the baseline model; similarly, when $\delta \geq \frac{(1-\alpha)u_y - u_x + \alpha u_p - \alpha(n-1)\beta}{(1-\alpha)(1-(1-\alpha)^{\hat{m}})(u_y - u_p) + (1-\alpha)(n-1)\beta - (1-\alpha)^{\hat{m}+1}\beta\hat{m} + (n-1)b - l(n-1-\hat{m})}$, the grim trigger strategy constitutes a subgame perfect equilibrium of the infinite repeated game of the main model.*

Proof of Proposition A.4. We prove that the grim trigger strategy is a Nash equilibrium, that is, players would not deviate from the full-cooperation outcome. For the baseline model, the non-deviation condition is given by:

$$\begin{aligned}\frac{u_x + (n-1)\beta - l(n-1)}{1-\delta} \geq\ & (1-\alpha)[u_y + (n-1)\beta] + \alpha u_p - l(n-1) \\ & + \frac{\delta}{1-\delta}[(1-\alpha)(u_y + (n-1)\beta) + \alpha u_p - (n-1)b - l(n-1)].\end{aligned}$$

Solving the inequality yields $\delta \geq \frac{(1-\alpha)u_y - u_x + \alpha u_p - \alpha(n-1)\beta}{(n-1)b}$. Similarly, the non-deviation condition for the main model is given by:

$$\begin{aligned}\frac{u_x + (n-1)\beta - l(n-1)}{1-\delta} \geq\ & (1-\alpha)[u_y + (n-1)\beta] + \alpha u_p - l(n-1) \\ & + \frac{\delta}{1-\delta}[(1-\alpha)^{\hat{m}+1}(u_y + \beta\hat{m}) + (1-(1-\alpha)^{\hat{m}+1})u_p - (n-1)b - l\hat{m}].\end{aligned}$$

Solving the inequality yields $\delta \geq \frac{(1-\alpha)u_y - u_x + \alpha u_p - \alpha(n-1)\beta}{(1-\alpha)(1-(1-\alpha)^{\hat{m}})(u_y - u_p) + (1-\alpha)(n-1)\beta - (1-\alpha)^{\hat{m}+1}\beta\hat{m} + (n-1)b - l(n-1-\hat{m})}$.

□

A. 3 Discussion: An alternative generalized model

In the generalized model presented in Appendix A, we assume that a player incurs a positive link cost only when forming a neighbor. In this section, we explore an alternative setting where a player pays a cost for each link proposal, regardless of whether a neighbor is formed or not, and briefly discuss the corresponding equilibrium properties.

In contrast to the assumption in Section 3, here we assume that the cost of each link proposal is positive ($c > 0$), irrespective of whether a neighbor is established or not. Let all other settings follow those in Section A. 1.1. Therefore, the payoff function is now represented as follows:

$$U_i(\mathbf{g}, \mathbf{a}) = U_{i,gain} - U_{i,loss} - c \sum_{j \in N, k \neq i} g_{ij}, \tag{12}$$

where $c \sum_{j \in N, k \neq i} g_{ij}$ is the total cost of link proposals for player i, and $U_{i,gain}$ and $U_{i,loss}$ remain as defined in Equations (7) and (8).

First, we present the equilibrium properties of the one-shot game for the baseline model in Proposition A.5 and the main model in Proposition A.6. In equilibrium, all players choose action y in both models. Moreover, no network with unilateral proposals can be supported in equilibrium in both models, as making unilateral proposals incurs a cost $c > 0$. Additionally, in the baseline model, players can form any number of neighbors in the equilibrium, while in the main model, players are limited to forming at most $\hat{m}$ neighbors. We omit a detailed proof of Proposition A.5 and A.6, as they follow the same steps as those in Section A. 2.

Proposition A.5. *Suppose $c \leq \beta(1-\alpha)$. For every subgame-perfect equilibrium of the baseline model, $(\boldsymbol{g}, \boldsymbol{a})$, the following properties hold for all $i, j \in N$ with $i \neq j$: (i) $g_{ij} = g_{ji}$ and (ii) $a_i = y$.*

Proposition A.6. *Let $\hat{m} = \arg\max_{m \in N}\{(1-\alpha)^{m+1}(u_y + m\beta) + [1-(1-\alpha)^{m+1}]u_p - cm\}$, and suppose $c \leq (1-\alpha)^{\hat{m}}\{ln(1-\alpha)[u_y + (\hat{m}-1)\beta - u_p] + \beta\}$. For every subgame-perfect equilibrium of the main model, $(\boldsymbol{g}, \boldsymbol{a})$, the following properties hold for all $i, j \in N$ with $i \neq j$: (i) $g_{ij} = g_{ji}$, (ii) $|N_i(\boldsymbol{g})| \leq \hat{m}$, and (iii) $a_i = y$.*

Next, we consider the refinement of elimination of weakly dominated strategies. However, unlike the scenarios discussed in Section 3.2.2, under the setting of costly unilateral link proposals, the strategy of proposing fewer links cannot be weakly dominated by the strategy of proposing more links. This is because a player would earn a lower payoff if they propose more links but fail to establish additional neighbors, compared to the strategy where they make fewer link proposals. As a result, the strategy of proposing fewer links cannot be weakly dominated by proposing more links, and consequently, this refinement does not apply in this setting.

Finally, we turn to the infinitely repeated game. Although the one-shot game exhibits multiple equilibria in both models, the noncooperation and full-cooperation equilibrium presented in Proposition 6 and Proposition 5 remain valid in this setting. This is because the one-shot strategies we use to construct strategies for the infinitely repeated game continue to be equilibria. We omit a detailed discussion of this here.

Appendix B Proofs

Proof of Proposition 1. We first prove that choosing action y is a strictly dominant strategy for any player i in the baseline model, given any network $\mathbf{g}$. Let $|N_i(\mathbf{g})|$ denote the number of neighbors of player i, and let M be the number of other players (excluding i) who choose action y. The expected payoff for player i when choosing action y is always strictly greater than the expected payoff of choosing action x for any network $\mathbf{g}$ and any M, regardless of the actions of other players:

$$0.85 \times (100 + 20|N_i(\mathbf{g})|) - 15M > 50 + 20|N_i(\mathbf{g})| - 15M.$$

Next, we prove that choosing action y is also a strictly dominant strategy for any player i in the main model, given any network $\mathbf{g}$. Let $|N_i(\mathbf{g})|$ again denote the number of neighbors of any player i, let m be the number of i's neighbors who choose y, and let M be the number of other players (excluding i) who choose action y. The expected payoff for player i when choosing action y is always strictly greater than expected payoff of choosing action x for any network $\mathbf{g}$, m, and M:

$$0.85^{m+1}(100 + 20 \times |N_i(\mathbf{g})|) - 15M > 0.85^m(50 + 20 \times |N_i(\mathbf{g})|) - 15M.$$

Therefore, all players choosing y is the unique equilibrium in the second-stage game for both the baseline and the main model. □

Proof of Proposition 2. In the baseline model, given any formed network and given that all players choose action y, players have no incentive to sever links with any established neighbors, as the link cost is zero and the benefit from each neighbor is positive. Additionally, players have no incentive to propose a link to others who are not neighbors, since the formation of neighbors requires mutual consent. Therefore, we only need to prove that no unilateral proposal exists in any equilibrium. We prove this by contradiction. Suppose, in an equilibrium, there exists two players $i \neq j \in N$, such that j proposes a link to i, but i does not propose to j (i.e., $g_{ij} = 0$ and $g_{ji} = 1$). In this case, player i can earn a greater expected payoff by proposing a link to j, as i would gain from having more neighbors, while the probability of being punished remains unchanged. □

Proof of Proposition 3. We first prove that in any equilibrium of the main model, each player has at most one neighbor. Given that all players choose y in the second stage, the expected payoff for a player with m neighbors, $0.85^{m+1}(100+20\times m)-15M$, is maximized when $m=1$. Therefore, if a player has more than one neighbor, they have an incentive to sever links and reduce their number of neighbors to one to earn a greater expected payoff.

Moreover, we prove that for any two players i and j, $g_{ji}=0$ if $|N_i(\mathbf{g})|=0$. Suppose $g_{ji}=1$, player i would have an incentive to propose a link to player j to earn a higher expected payoff than by having no neighbor. □

Proof of Proposition 4. First, we prove that the strategy in which player i proposes $k<7$ links, and chooses action y, denoted as $s_i^k=(\mathbf{g}_i,y)$, where $g_{ij}\in\{0,1\}$ for each player $j\neq i$ and $\sum_{j\neq i} g_{ij}=k$, is weakly dominated by the strategy in which the player proposes links to all other players and chooses y, denoted as $s_i^7=(\mathbf{g}_i,y)$, where $g_{ij}=1$ for each player $j\neq i$. Here, the superscript of s_i^k and s_i^7 refers to the number of link proposals. Specifically, given any profile of other players' strategies, let m denote the number of neighbors that player i establishes when adopting strategy s_i^k, and let $\bar{m}$ denote the number of neighbors for player i when adopting strategy s_i^7. Notably, the condition $\bar{m}\geq m$ always holds because $k<7$. Furthermore, $\bar{m}>m$ occurs when each of the other seven players adopts the strategy of proposing links to all players and choosing y, which results in $\bar{m}=7$ while $m<7$. Consequently, given any profile of other players' strategies, the expected utility of player i adopting s_i^7 is at least as high as that of s_i due to the following inequality:

$$0.85\times(100+20\bar{m})-15\times M\geq 0.85\times(100+20m)-15\times M,$$

where M is the number of other players in the group who choose y. This inequality holds strictly when $\bar{m}>m$. Therefore, any equilibrium in which some players adopt s_i^k cannot survive the elimination of weakly dominated strategies. Conversely, s_i^7 cannot be weakly dominated by any strategy. Hence, all players adopting s_i^7 is the unique equilibrium that survives the elimination of weakly dominated strategies.

For the main model, we demonstrate that the strategy in which player i proposes zero links and chooses action y, denoted as $s_i^0=(\mathbf{g}_i,y)$, where $g_{ij}=0$ for each player $j\neq i$, is weakly dominated by the strategy in which the player proposes one link and chooses

y, denoted as $s_i^1 = (\mathbf{g}_i, y)$, where $\sum_{j \neq i} g_{ij} = 1$. The logic behind this proof mirrors that applied in the baseline model. Notably, when adopting s_i^1, player i either establishes one neighbor or zero neighbors, depending on the profile of other players' strategies. In contrast, when adopting s_i^0, player i consistently establishes zero neighbors. If player i establishes one neighbor by adopting s_i^1, the expected payoff from s_i^1 is strictly higher than that from s_i^0, regardless of the action choice made by i's neighbor. This is evident from the following inequalities:

$$0.85^2 \times (100 + 20) - 15 \times M > 0.85 \times 100 - 15 \times M,$$

and

$$0.85 \times (100 + 20) - 15 \times M > 0.85 \times 100 - 15 \times M,$$

where, in the first inequality, the neighbor chooses y, and in the second inequality, the neighbor chooses x, and M is the number of other players in the group who choose y. On the other hand, if player i establishes zero neighbors by adopting s_i^1, the expected payoff remains the same between adopting s_i^1 and s_i^0. Therefore, the expected payoff of adopting s_i^1 is at least as high as that of adopting s_i^0. Consequently, the strategy of s_i^0 is weakly dominated by s_i^1, and thus, the equilibrium in which some players propose zero links and choose y cannot survive the elimination of weakly dominated strategy in the main model. Importantly, none of the other equilibrium strategies in the main model can be weakly dominated by any other strategies. □

Proof of Proposition 6. We prove that the grim trigger strategy constitutes a subgame-perfect equilibrium of the infinitely repeated game. To demonstrate this, it suffices to show that the grim trigger strategy is a Nash equilibrium, as it is inherently subgame-perfect: after any deviation from the full-cooperation outcome, players revert to an equilibrium of the one-shot game in both models. We first consider the baseline model. Each player receives a payoff of 190 per period from the full-cooperation outcome. If a player deviates by choosing y in some period t, their deviation payoff is $204 + \frac{\delta}{1-\delta}[(20*7+100)*0.85-15*7)]$. The condition for no deviation is:

$$\frac{190}{1-\delta} \geq 204 + \frac{\delta}{1-\delta}[(20 * 7 + 100) * 0.85 - 15 * 7)].$$

Solving this inequality yields $\delta \geq 0.13$. Similarly, in the main model, each player receives a payoff of 190 per period from the full-cooperation outcome. If a player deviates by choosing y in some period t, their deviation payoff is $204 + \frac{\delta}{1-\delta}[(20+100) * 0.85^2 - 15 * 7]$. The condition for no deviation is:

$$\frac{190}{1-\delta} \geq 204 + \frac{\delta}{1-\delta}[(20+100) * 0.85^2 - 15 * 7].$$

Solving this inequality yields $\delta \geq 0.06$. □

Proof of Proposition 7. We first provide the calculation for the size of the basin of attraction respectively for the baseline and the main model.

Suppose a player assigns a probability p to each other player in the group adopting the grim trigger strategy (denoted as G), and $1-p$ to each other player adopting the noncooperation strategy as outlined in Proposition 4 (denoted as NC). The size of the basin of attraction of NC, *sizeBNC*, corresponds to the probability that equalizes the expected payoff associated with either strategy.

We first calculate the size of the basin of attraction of NC for the baseline model, denoted as $sizeBNC^B$. The probability that there are in total m players among all other seven group members choosing G is $C_7^m p^m (1-p)^{7-m}$. In the infinitely repeated game, NC results in an expected payoff of

$$\begin{aligned}\sum_{m=0}^{7} C_7^m p^m (1-p)^{7-m} \{&[0.85 \times (100 + 7 \times 20) - 15 \times (7-m)] \\ &+ \frac{\delta}{1-\delta}[0.85 \times (100 + 20 \times (7-m)) - 7 \times 15]\},\end{aligned} \tag{13}$$

and G results in an expected payoff of

$$\begin{aligned}p^7 \frac{1}{1-\delta}(50 + 7 \times 20) + \sum_{m=0}^{6} C_7^m p^m (1-p)^{7-m} \{&[(50 + 7 \times 20) - 15 \times (7-m)] \\ &+ \frac{\delta}{1-\delta}[100 \times 0.85 - 7 \times 15]\}.\end{aligned} \tag{14}$$

Therefore, $sizeBNC^B$ equalizes the formula (13) and (14):

$$\begin{aligned}&\sum_{m=0}^{7} C_7^m p^m (1-p)^{7-m} [99 + 15m + \frac{\delta}{1-\delta}(99 - 17m)] \\ &= p^7 \frac{1}{1-\delta} 190 + \sum_{m=0}^{6} C_7^m p^m (1-p)^{7-m} [85 + 15m - \frac{\delta}{1-\delta} 20].\end{aligned} \tag{15}$$

Next we solve the *sizeBNC* of the mian model, denoted as $sizeBNC^M$. The strategy of NC requires players in every period propose a single link and form a 1-regular network where everyone has a single neighbor. Denote the neighbor of player i in NC as player j.

The probability that there are m players who adopt strategy G among other six group members (excluding i and j) is $C_6^m p^m(1-p)^{6-m}$. In the infinitely repeated game, NC results in an expected payoff of $\sum_{m=0}^{6} C_6^m p^m(1-p)^{6-m}\{[0.85^2\times(100+20)-15\times(7-m)]+\frac{\delta}{1-\delta}[0.85^2(100+20)-7\times 15]\}$ against the player j following NC, or an expected payoff of $\sum_{m=0}^{6} C_6^m p^m(1-p)^{6-m}\{[0.85\times(100+20)-15\times(6-m)]+\frac{\delta}{1-\delta}[0.85\times 100-7\times 15]\}$ against the player j following G. Overall, NC yields an expected payoff of

$$\begin{aligned}(1-p)*\sum_{m=0}^{6} C_6^m p^m(1-p)^{6-m}\{[0.85^2\times(100+20)-15\times(7-m)]\\ +\frac{\delta}{1-\delta}[0.85^2(100+20)-7\times 15]\}\\ +p*\sum_{m=0}^{6} C_6^m p^m(1-p)^{6-m}\{[0.85\times(100+20)-15\times(6-m)]\\ +\frac{\delta}{1-\delta}[0.85\times 100-7\times 15]\}.\end{aligned} \tag{16}$$

Similarly, G results in an expected payoff of $\sum_{m=0}^{6} C_6^m p^m(1-p)^{6-m}\{0.85\times(50+20(m+1))-15\times(7-m)+\frac{\delta}{1-\delta}[0.85\times 100-7\times 15]\}$ against the player j following NC, an expected payoff of $\sum_{m=0}^{5} C_6^m p^m(1-p)^{6-m}\{(50+20(m+1))-15\times(6-m)+\frac{\delta}{1-\delta}[0.85\times 100-7\times 15]\}$ against the player j following G and at least one player in the group following NG, or an expected payoff of $(50+7\times 20)\frac{1}{1-\delta}$ against all other participants following G. Thus G yields an expected payoff of

$$\begin{aligned}p^7(50+7\times 20)\frac{1}{1-\delta}\\ +(1-p)*\sum_{m=0}^{6} C_6^m p^m(1-p)^{6-m}\{0.85\times(50+20(m+1))-15\times(7-m)\\ +\frac{\delta}{1-\delta}[0.85\times 100-7\times 15]\}\\ +p*\sum_{m=0}^{5} C_6^m p^m(1-p)^{6-m}\{(50+20(m+1))-15\times(6-m)\\ +\frac{\delta}{1-\delta}[0.85\times 100-7\times 15]\}.\end{aligned} \tag{17}$$

Consequently, $sizeBNC^M$ equalizes the formula (16) and (17):

$$\begin{aligned}
&(1-p)*\sum_{m=0}^{6} C_6^m p^m (1-p)^{6-m}[15m - 18.3 - \frac{\delta}{1-\delta}18.3] \\
&\quad +p*\sum_{m=0}^{6} C_6^m p^m (1-p)^{6-m}[12 + 15m - \frac{\delta}{1-\delta}20] \\
=\, & p^7\frac{1}{1-\delta}190 + (1-p)*\sum_{m=0}^{6} C_6^m p^m (1-p)^{6-m}[32m - 45.5 - \frac{\delta}{1-\delta}20] \\
&\quad +p*\sum_{m=0}^{5} C_6^m p^m (1-p)^{6-m}[35m - 20 - \frac{\delta}{1-\delta}20].
\end{aligned} \tag{18}$$

Equation (15) and (18) can be respectively reorganized as

$$14 + \frac{\delta}{1-\delta}119(1-p) - p^7\frac{\delta}{1-\delta}210 = 0, \tag{19}$$

and

$$(1-p)(1.7\frac{\delta}{1-\delta} + 27.2 - 102p) + (32 - 120p)p - p^7\frac{\delta}{1-\delta}210 = 0. \tag{20}$$

Note that the left-hand sides of Equation (19) and (20) both decrease in p given $0 \leq p \leq 1$. If the left-hand side of (19) is always greater than that of (20) given any $p \in [0, 1]$, then $sizeBNC^M$ must be smaller than $sizeBNC^B$. A sufficient condition for this to hold true is $\delta \geq 0.55$. □

Appendix C Experimental Instructions

This appendix is an English translation of the instructions for the treatment CP. Treatment specific texts for RS and PI are shown in italics.

Welcome

Welcome to this experiment on decision-making. Please read the following instructions carefully. During the experiment, do not communicate with other participants in any means. If you have any questions at any time, please raise your hand, and an experimenter will come to assist you privately. This experiment will last about 90 minutes. You will participate in an experiment in this room with other participants, each seated behind a separate computer, without knowledge of each other's identity. All decisions will be made anonymously on the computer screen. Experimenters and other participants will not be able to link your name to your desk number or to the decisions you make.

Your earnings in this experiment will be measured in points. Your earnings depend on your own choices and the choices of other participants. At the end of the experiment, your earnings will be converted to Chinese *yuan* at the rate of 80 points to 1 *yuan*. You will also receive a show-up fee of 15 *yuan* which will be added to your earnings. Your earnings will be paid to you privately.

Please read the following instructions carefully. Afterward, we will ask you several questions to ensure your understanding. The will not proceed until all participants answer the questions correctly.

In this experiment, you will be randomly matched with seven other participants in the room. Your group of eight will remain unchanged throughout the experiment. Each of the eight participants will be assigned a player ID, which will remain constant throughout the experiment and is one of the following: A, B, C, D, E, F, G and H. The eight of you are going to repeatedly play a game for several rounds. The game contains two [*PI* & *RS: three*] decision stages.

The first-stage decision: Neighbors formation

In the first stage, the eight of you will simultaneously decide whether to propose interaction to other seven players respectively. You can make any proposals as you want. (You can propose interaction to all players, and you can also decide not to make any proposal.) If you make a proposal to one player, it is a **one-way interaction**. If this player also proposes to interact with you, the two of you can form a **two-way interaction**. As the decisions are made simultaneously, no player knows whether others interaction with them when they make decisions. Specifically,

- If two players both propose interaction to each other, they form the two-way interaction. We call the players who form the two-way interaction with you as your **neighbors**. That is to say, **mutual consent** is needed for two players to establish a neighbor relationship.
- Two players cannot not form the two-way interaction and therefore are not neighbors if at least one of them does not propose interaction to the other.

The result of interaction proposals will be displayed as a network graph to everyone after all players in the group finish making proposals in the first stage. Each player is represented as a gray dot marked with the identity ID (A, B, C, D, E, F, G and H). A **thick complete line** between two players indicates they have formed a two-way interaction and are neighbors. A **thin incomplete line** between two players indicates that only one of them proposed interaction and therefore they are not neighbors. The line starts from the side of the person who proposed interaction and ends just before the dot of the person who did not propose. If two players did not propose interaction to each other, there will be no line between them.

The second-stage decision I: Action choices, earnings and losses

In the second stage, you can first observe the network formed in the first stage, then you have to choose between two actions: X and Y. Your payoffs can be separated into two components: payoff = earning - loss.

The earning contains two parts, basic earning and interaction earning.

- **Basic earning**: The basic earning of action X is 50 points, and the basic earning of action Y is 100 points.
- **Interaction earning**: In addition to the basic earning from actions, you can also earn through interaction with more neighbors. Interaction earning will depend on the number of neighbors you have, and you can earn 20 points for each additional neighbor. In other words, if you have k neighbors, your interaction earning will be $20 * k$. (Please note that a neighbor is defined as a player who has formed a two-way interaction with you. You cannot earn from players with you only have a one-way interaction.)

The loss is computed as follows:

- **Loss**: Each player selecting action Y will result in a loss of 15 points for each of the seven other players in the group. However, if you choose action Y, it will not lead to a loss for yourself. Hence, if there are n players in your group who choose Y, your loss will be $15 * n$. By the same token, when you choose Y, you will cause the other seven players to suffer a loss of 15 points each (thus 70 points in total).

In summary, your payoff is:

- If you choose X, your payoff = $50 + 20 * k$ (the number of you neighbors)- $15 * n$ (the number of action Y chosen by the other seven players)

- If you choose Y, your payoff = $100 + 20 * k$ (the number of you neighbors)- $15 * n$ (the number of action Y chosen by the other seven players)

The second-stage decision II: Spot check and punishment

After all players finish choosing actions, the computer randomly checks each player's choice with a probability of 15%. You will be punished if you or your neighbors are found to have chosen Y during the spot check. Each player has an equal probability of being checked (15%) in each round, and whether or not they are checked is an independent event, unaffected by whether their neighbors are checked. (For instance, if Player A is checked, it does not guarantee that players B-H will be checked or not, and neither is it a guarantee that A's neighbors are checked or not. The number of checked players per round can vary between 0 to 8.)

You are punished if and only if you choose action Y and you are checked by the computer. If you are punished, your earning will be made zero.

[*RS & PI: If you and one of your neighbors are checked and found to have chosen Y, your earning will be made zero. To be specific,*

- *If at least one of you and your neighbors are checked and found to have chosen action Y, you will be punished, and the earning is 0 in this round.*
- *If all of you and your neighbors are not checked, or are checked but found to have chosen action X, your earning is not changed.*]

Please note that the punishment will lead to zero earnings but will not impact the loss component of your payoff. The table outlines the associated points for each scenario:

	You choose X	You choose Y
If you are not checked [*RS & PI: If you and all your neighbors are not found to have chosen Y in the spot check*]	50 + 20*The number of neighbors - 15*The number of action Y chosen by the other seven players	100 + 20* The number of neighbors - 15* The number of action Y chosen by the other seven players
If you are checked [*RS & PI: If at least one of you and your neighbors are found to have chosen Y in the spot check*]	0-15* The number of action Y chosen by the other seven players	

There is a page pf paper on your desk displaying the aforementioned payoff table, which you can consult during the experiment. [*RS & PI: Therefore, the probability of being punished directly depends on you and your neighbors action choices. There is a page of paper on your desk, in which we list the probability of being punished given different number of neighbors choosing Y, as well as the aforementioned payoff table which you can consult during the experiment.*]

The number of Y chosen by you and your neighbors	1	2	3	4	5	6	7	8
The probability that at least one of you and your neighbors are found to have chosen Y	15%	28%	39%	48%	56%	62%	68%	73%

*(Note: This table pertains exclusively to RS and PI treatments.)

[*PI: The third-stage decision: Disclosing action choices*

In this stage, all players must independently decide whether to disclose their action choice (X or Y) to the entire group at the beginning of the next round. Opting for 'Disclosure' means that, during the first stage of the next round (the neighbor formation stage), all group members will be able to view your action choice and payoff in the previous round. Choosing 'Not disclosing' means that other players will be unable to view your action choice and payoff. It's important to note, however, if you are found to have chosen Y during the spot check in this round, your actions will be disclosed to other players, regardless of whether you choose to disclose or not.

Therefore, at the beginning of each round, you can observe the action choices and payoffs of players who decide to disclose, as well as those whose action Y are founded in the previous round. You will not be able to observe the action choices or payoffs of those who choose not to disclose.

Please note that the neighbor network is public information in each round, and you can observe link proposals from all players, regardless of their disclosure decisions.

The cost of disclosing is 5 points each round, and it will be deducted from your next round's payoff since actions not be released until then. If you choose not to disclose, there *is no cost. In summary, the final payoff for each round is as follows:*

- *If you choose disclosing in the previous round, your final payoff = the second-stage payoff - 5.*
- *If you choose disclosing in the previous round, your final payoff = the second-stage payoff.*]

The ending rule of the experiment

You will be playing this game with the same group of individuals for several rounds. The number of repetition rounds is randomly determined by computer. At the end of each round, there is a 3% probability that the computer will end the experiment, while there is a 97% probability that all eight individuals will proceed to the next round. Regardless of the round number, the probability of proceeding to the next round remains fixed at 97%. The repetition rounds are the same for everyone in group.

History information and operation instructions

You will have an unpaid trial period to become familiar with the experiment. After that, at the start of each round, you will be able to view all past outcomes. The node of any player who has been checked will be displayed in green. [*PI: For players who choose not to disclose (except for players who are found choosing Y), their actions and payoffs will be displayed as blank.*] In every stage of the experiment, you can click on each player's node to see all interaction proposals of that player (both those proposed by him and those proposing to him), and click again to return the overall network graph.

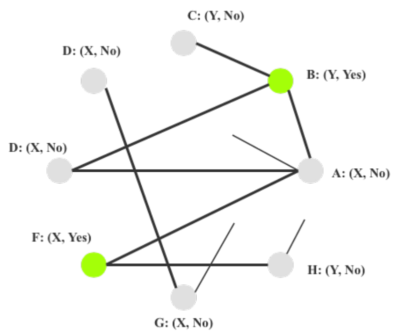
Incomes

The total points you earned are computed as the sum of points in each round. All points will be converted to Chinese *yuan* at the rate: 80 points = 1 *yuan*. In addition, you get a show-up fee of 15 *yuan*. Your income in the experiment is therefore computed as follows:

$$\text{Income} = 15 + \text{total points} /80$$

Control questions

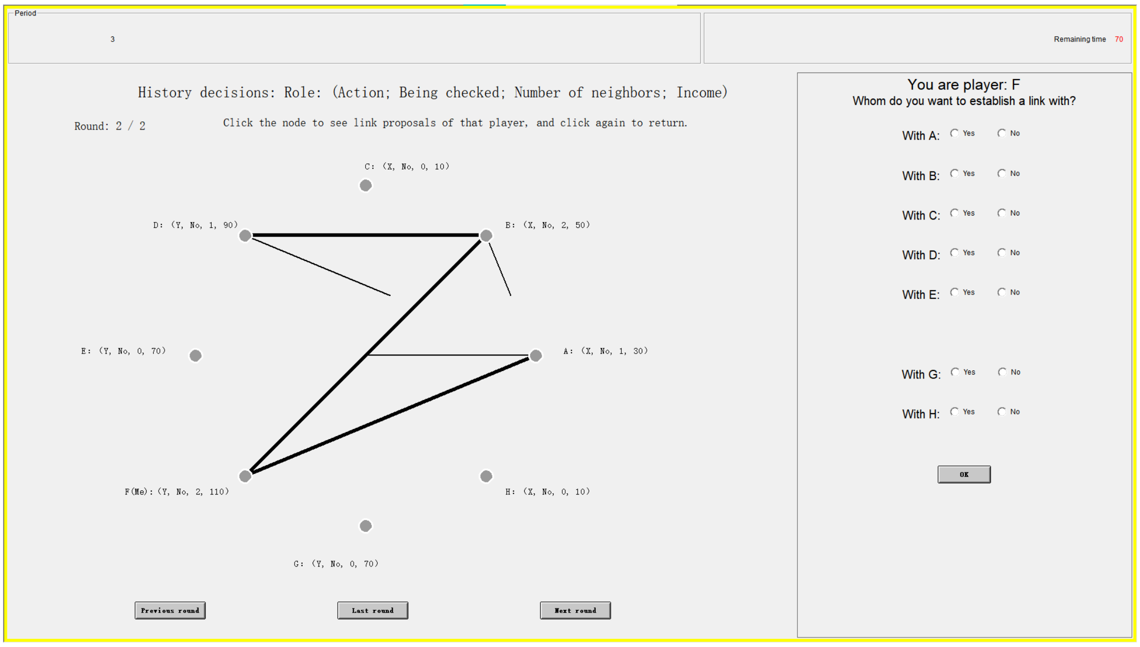
1. Suppose in one round, players form a network displayed below (with actions and c outcomes shown in parenthesis). Who are the neighbors of player A?



2. Who are checked in this round?
3. What is player A's earning?
4. What is player B's earning?
5. What is player F's earning?
6. What is player H's earning?
7. Which of the following statements is correct? (i) I will play with the same group of individuals in every round; (ii) I will never play with the same group of individuals for more than one round; (iii) I might play with the same group of individuals for more than one round.
8. Which of the following statements is correct? (i) As the game progresses through more rounds, the probability of ending the game at the current round increases; (ii) regardless of the round the game has reached, the probability of ending the game at the current round is always 3%.
9. Which of the following statements is correct? (i) Regardless of whether I choose action X or action Y, my probability of being checked is 15%. (ii) The more neighbors I have who choose action Y, the higher my probability of being checked; (iii) The more neighbors I have who choose action Y, the lower my probability of being checked.
10. If I am not checked for choosing action Y, which of the following situations may occur? (i) I chose action X, and I am not checked; (ii) I chose action X, and I am checked; (iii) I chose action Y, and I am not checked.

Screenshots

Figure 4: Screenshot of the first-stage decision



Period

3

Remaining time 70

History decisions: Role: (Action; Being checked; Number of neighbors; Income)

Round: 2 / 2

Click the node to see link proposals of that player, and click again to return.

You are player: F

Whom do you want to establish a link with?

With A: ○ Yes ○ No

With B: ○ Yes ○ No

With C: ○ Yes ○ No

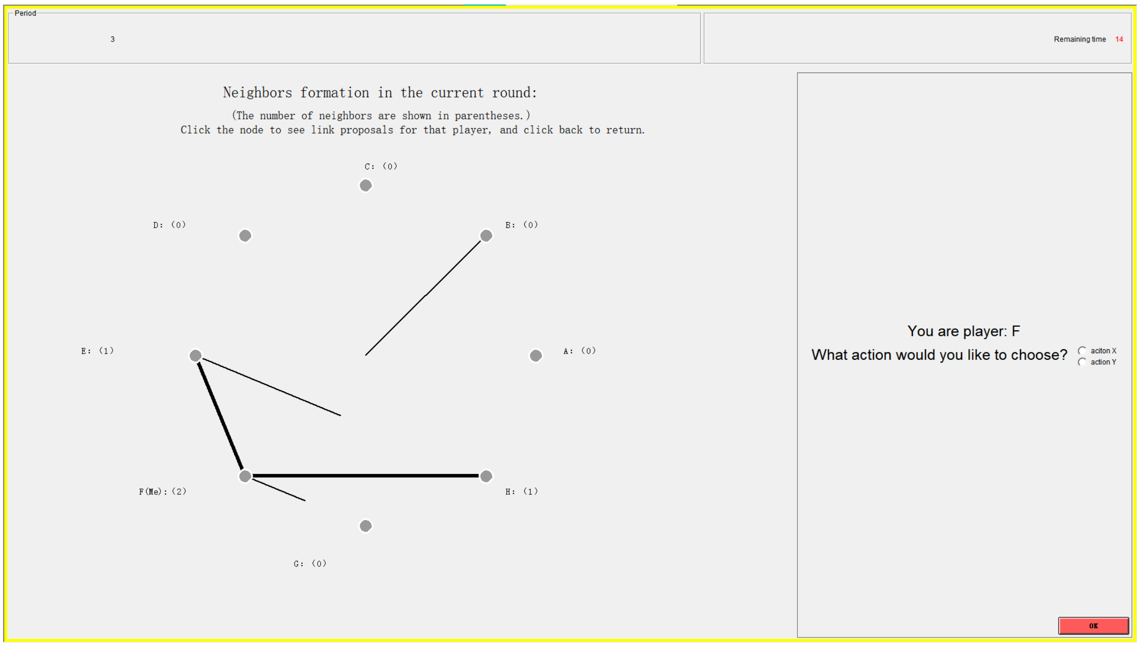
With D: ○ Yes ○ No

With E: ○ Yes ○ No

With G: ○ Yes ○ No

With H: ○ Yes ○ No

Figure 5: Screenshot of the second-stage decision



Period

3

Remaining time 14

Neighbors formation in the current round:

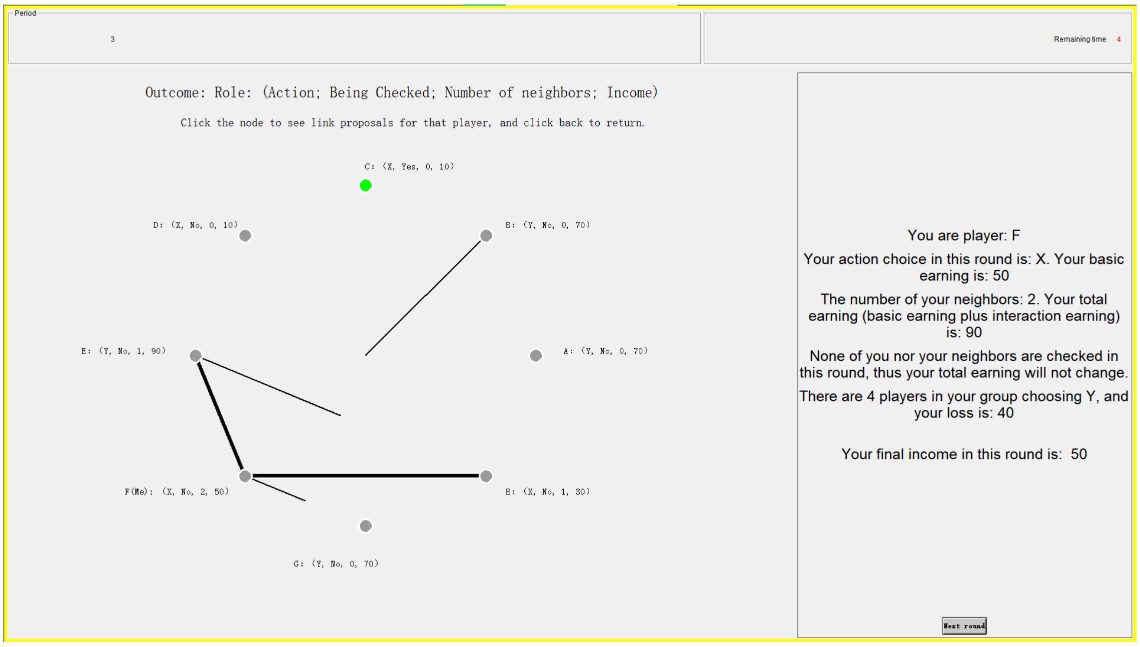
(The number of neighbors are shown in parentheses.)

Click the node to see link proposals for that player, and click back to return.

You are player: F

What action would you like to choose? ○ aciton X ○ action Y

Figure 6: Screenshot of the outcome presentation stage



Outcome: Role: (Action; Being Checked; Number of neighbors; Income)

Click the node to see link proposals for that player, and click back to return.

You are player: F

Your action choice in this round is: X. Your basic earning is: 50

The number of your neighbors: 2. Your total earning (basic earning plus interaction earning) is: 90

None of you nor your neighbors are checked in this round, thus your total earning will not change.

There are 4 players in your group choosing Y, and your loss is: 40

Your final income in this round is: 50

Appendix D Supplemental Tables and Figures

Table 5: Extended statistics of cooperation levels and number of neighbors

	Average cooperation level			Average number of neighbors		
Rounds	21-30	21-40	21-50	21-30	21-40	21-50
CP	2.87	2.89	2.89	6.90	6.91	6.91
	(1.13)	(1.07)	(1.08)	(0.21)	(0.20)	(0.20)
RS	7.77	7.74	7.75	6.69	6.71	6.69
	(0.45)	(0.56)	(0.56)	(0.40)	(0.36)	(0.37)
PI	4.36	4.37	4.36	5.16	5.12	5.10
	(3.42)	(3.39)	(3.41)	(1.70)	(1.74)	(1.77)
	Mann-Whitney test (p-value)					
CP v.s. RS	<0.001	<0.001	<0.001	0.339	0.339	0.233
CP v.s. PI	0.870	0.966	0.966	0.054	0.054	0.054
PI v.s. RS	0.078	0.106	0.106	0.143	0.143	0.196

Notes: Each cell shows the average number of participants choosing action x and the average degree for rounds 21-30, 21-40 or 21-50 (standard deviations in parentheses). The two-sided Mann-Whitney test is performed at the group level and $n = 14$.

Table 6: Robustness of determinants of participant i's public bad actions

	(1) CP	(2) RS	(3) PI	(4) PI-C	(5) PI-NC
$\beta_1 : y_{i,t-1}$	0.712***	1.277***	1.887***	0.728	1.905***
	(0.126)	(0.341)	(0.247)	(0.759)	(0.275)
$\beta_2 : \%cooperation_{j\neq i,t-1}$	-0.203***	-0.534***	-0.294***	-0.168	-0.281***
	(0.050)	(0.080)	(0.084)	(0.205)	(0.095)
$\beta_3 : \#neighbors_{i,t}$	0.771***	-0.320***	-0.369***	-0.785***	-0.299***
	(0.205)	(0.086)	(0.063)	(0.172)	(0.074)
$\beta_4 : \#failed_{i,t}$	2.280***	-0.085	-0.042	-0.273	-0.018
	(0.596)	(0.133)	(0.119)	(0.246)	(0.137)
$\beta_5 : y_{i,t-1}\times\ \#failed_{i,t}$	-1.556***	-0.090	-0.043	-0.116	0.089
	(0.601)	(0.125)	(0.132)	(0.248)	(0.160)
$\beta_6 : detection_{i,t-1}$	1.381***	0.919***	-0.594	0.174	-1.025*
	(0.248)	(0.323)	(0.453)	(0.748)	(0.557)
$\beta_7 : punished^{detection}_{i,t-1}$	-0.411	0.053	-0.212	1.014	-0.383
	(0.352)	(0.595)	(0.539)	(1.122)	(0.640)
$\beta_8 : punished^{neighbor}_{i,t-1}$		-0.005	0.250	0.699	0.079
		(0.384)	(0.226)	(0.747)	(0.240)
N	1979	1224	1400	487	913

The reported results are parameter estimates from fixed-effects logit regressions.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The main difference between Table 6 and Table 3 is that the former employs a fixed-effects logit panel regression. Due to the issue of insufficient variation, approximately half of the observations were dropped. Nevertheless, most of the results in Table 3 remain robust. The only difference is that, in PI-C, β_1 and β_2 are no longer statistically significant, which may be attributed to the reduction in sample size.

Table 7: Determinants of participant i's disclosure for both actions

	(1) PI	(2) PI-C	(3) PI-NC
$\beta_1 : y_{i,t}$	-0.279*** (0.028)	-0.483*** (0.038)	-0.201*** (0.013)
$\beta_2 : d_{i,t-1}$	0.066** (0.031)	0.132*** (0.043)	0.000 (0.024)
$\beta_3 : \%d_{t-1}$	0.172* (0.102)	0.093 (0.171)	0.265*** (0.065)
$\beta_4 : \#failed_{i,t}$	0.021*** (0.007)	0.028 (0.017)	0.014** (0.007)
$\beta_5 : \#neighbors_{i,t}$	0.003 (0.006)	0.002 (0.014)	0.005 (0.007)
$\beta_6 : lowcoop_t$	-0.028 (0.043)	-0.181*** (0.024)	0.019 (0.015)
$\beta_7 : highcoop_t$		(benchmark)	
$\beta_8 : fullcoop_t$	-0.218*** (0.049)	-0.275*** (0.083)	
$\beta_9 : y_{i,t-1}$	0.054 (0.036)	0.202*** (0.075)	0.005 (0.030)
N	2136	1016	1120

Standard errors (in parentheses) are clustered at the group level. The reported results are marginal effects from random effects probit regressions.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Robustness of determinants of participant i's disclosure of action x

	(1) PI	(2) PI-C	(3) PI-NC
$\beta_1 : d_{i,t-1}$	0.996***	0.972***	1.055**
	(0.252)	(0.291)	(0.512)
$\beta_2 : \%d_{t-1}$	2.340***	2.183***	3.512**
	(0.565)	(0.614)	(1.587)
$\beta_3 : \#failed_{i,t}$	0.465***	0.458***	0.415***
	(0.092)	(0.132)	(0.137)
$\beta_4 : \#neighbors_{i,t}$	0.262***	0.283***	0.347***
	(0.074)	(0.109)	(0.122)
$\beta_5 : lowcoop_t$	-1.514***	-2.351*	-1.112**
	(0.390)	(1.268)	(0.447)
$\beta_6 : highcoop_t$		(benchmark)	
$\beta_7 : fullcoop_t$	-3.537***	-3.605***	
	(0.450)	(0.480)	
$\beta_8 : y_{i,t-1}$	-0.439	0.055	-0.617
	(0.304)	(0.497)	(0.481)
N	1632	974	658

The reported results are parameter estimates from fixed effects logit regressions.

A subset of observations is dropped due to insufficient variation.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The main difference between Table 8 and Table 4 is that the former employs a fixed-effects logit panel regression. The results in Table 4 remain robust. Compared with participants in PI-NC, those in PI-C respond more positively—in terms of coefficient magnitude—to changes in the previous round's disclosure rate of x (β_2), the number of failed proposals in the current round (β_3), and the number of neighbors in the current round (β_4). Furthermore, the effect of the current cooperation level on disclosure willingness continues to exhibit a negative-to-positive pattern: in PI-C, participants exhibit a stronger willingness to disclose at an earlier stage as the cooperation level increases (β_7).