[a]IAER, Dongbei University of Finance and Economics

创业与管理学院

School of Entrepreneurship and Management

SHANGHAITECH SEM WORKING PAPER SERIES
No. 2021-009

Reciprocity with Uncertainty About Others

Jin-yeong Sohn

IAER, Dongbei University of Finance and Economics

Wenhao Wu

SEM, ShanghaiTech University

September 2021

https://ssrn.com/abstract=3926177

School of Entrepreneurship and Management

ShanghaiTech University

http://sem.shanghaitech.edu.cn

Reciprocity with Uncertainty About Others*

Jin-yeong Sohn[a], Wenhao Wu[b,*]

[a]IAER, Dongbei University of Finance and Economics, 217 Jianshan St, Dalian, Liaoning, 116025 China

[b]SEM, ShanghaiTech University, 393 Middle Huaxia Road, Shanghai, 201210 China

Abstract

We introduce the uncertainty of psychological motivation into a reciprocity model and explore its implications on behavior. We extend the Sequential Reciprocity Equilibrium in extensive-form games (Dufwenberg and Kirchsteiger, 2004) to a broader class of incomplete information games. We use this general framework to study many well-known games. We investigate how uncertainty changes the equilibrium predictions of the standard reciprocity model and compare two setups in which the psychological motivations are known and unknown among the players, respectively. We find that, in the prisoners' dilemma, players are more likely to cooperate with each other when they have information about the reciprocal motivations of their opponents, given that the benefits of defection are modest.

JEL Classification: A13, D63, D81, D91

Keywords: Social Preferences, Reciprocity, Incomplete Information, Prisoners' Dilemma

1. Introduction

Behavioral economists have noticed that uncertainty about people's psychological motivations prevails. Attanasi, Battigalli and Manzoni (2016) have argued that it is implausible to assume that subjects who are randomly drawn from a population to participate in an experiment would have sufficient information to know their fellow subjects' other-regarding preferences. Furthermore, various experiments have suggested there

*We are grateful to Martin Dufwenberg for his invaluable comments and encouragement. We also thank Andreas Blume, Rachel Mannahan, Asaf Plan, and seminar attendees at the University of Arizona and at the 2019 Korean Econometric Society Conference for helpful comments.

* Corresponding author

is considerable heterogeneity among individuals' social motives. For example, Dohmen, Falk, Huffman and Sunde (2008) present survey results which demonstrate that individuals display a large degree of heterogeneity in trust and reciprocity. Hennig-Schmidt, Sadrieh and Rockenbach (2010) report a significant degree of heterogeneity in employees' effort levels after the employer offered them a bonus, which suggests that they feel grateful to the employer to different extents. Bellemare, Sebald and Suetens (2018) have also provided evidence that, in the dictator game, dictators exhibit varying levels of sensitivity to guilt. These laboratory experiments suggest that in many situations it may not be innocuous to assume complete information about psychological motivations.

There have been papers that discuss the implications of assuming incomplete information in psychological games. For instance, Battigalli and Dufwenberg (2009) lay out a general foundation for dynamic psychological games and they also extend the analysis to incomplete information games. Attanasi, Battigalli and Manzoni (2016) study guilt aversion in Bayesian games and explore how incomplete information influences the analyses in centipede games. Following their work, we contribute to this literature by focusing on *reciprocity games* in which players are uncertain about the intensity of each other's psychological motivations.

Previous reciprocity models investigate people's natural tendency to reward kind people and punish mean ones, and incorporate this tendency into standard game theory (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010; Antler, 2015). But they maintain the assumption of complete information about reciprocal motivations. We extend the reciprocity model in extensive-form games (Dufwenberg and Kirchsteiger, 2004; henceforth DK) to incorporate incomplete information about sensitivity parameters, which indicate the degrees to which players care about social motives. In other words, we assume the sensitivity parameters are unknown *types*, the realizations of which constitute private information for each player.

Due to information asymmetry, players do not know the types of each other and hence do not know the corresponding actions the other players will take. Then, whatever decision the player makes, there may be an unavoidable risk that he might *be kind to the unkind and be unkind to the kind* from the *ex post* perspective. This is a new strategic consideration that arises in the reciprocity game of incomplete information. Therefore, we evaluate a player's kindness toward others with respect to the consequence of the player's action on the

expectation of other players' well-being.

In Section 3, we apply our theory to a series of well-known games[2] and find that the direction of the effect of uncertainty on equilibrium outcomes is ambiguous. It depends on whether in equilibrium, (i) all types take the same action; or (ii) actions vary across types. In case (i), the equilibrium outcome may resemble that of a complete information game. For example, reciprocity has been used as an explanation for positive offers in ultimatum games (Van Damme et al., 2014). In Section 3.2, we analyze an ultimatum game in which the responder could be either reciprocal or selfish, and the probability of being reciprocal is over one-half. Then, without knowing the responder's type, the proposer will make a high offer as if the responder were reciprocal, despite the fact that the responder has a chance to be selfish and would like to accept the lowest offer. The similar pattern is also found in monopoly pricing (Sections 3.3). On the contrary, in case (ii), introducing uncertainty can yield entirely different predictions from the reciprocity model of complete information. In the battle of the sexes, for example, the strategy profile, in which both players choose the options the other players prefer, can be supported as a reciprocal equilibrium by negative reciprocity (Rabin, 1993). However, in Section 3.6, we show that this equilibrium may not be robust to an arbitrarily small perturbation of type distributions, in the sense that the possibility that each player chooses the other player's preferred action would be completely precluded.

One innovation of our analyses is that we make comparisons between *acquaintance* and *stranger* societies, where the types are known and unknown, respectively. We investigate whether uncertainty strengthens or weakens certain reciprocity equilibrium outcomes in the stranger society, i.e., whether it increases or decreases the *ex ante* probability of achieving those outcomes. In the sequential prisoners' dilemma (Section 3.4), for instance, suppose the second player could be selfish or reciprocal, but is more likely to be reciprocal. Then, the selfish type of the second player would defect regardless of whether the first player cooperates or not. But the reciprocal type is more willing to cooperate conditional on the first player's choosing cooperation in the stranger society, because she recognizes that the first player risks being failed by a selfish type when making this decision. Therefore, the first player is more likely to also cooperate rather than end up with mutual defection, and the probability of cooperation is higher in the stranger society than in the

[2]Our applications include the investment game, the ultimatum game, monopoly pricing, the sequential prisoners' dilemma, the public goods game, the battle of the sexes, and the prisoners' dilemma.

acquaintance society. In a public goods game, the expected contribution level could also be higher in the stranger society (Section 3.5). This is because in the acquaintance society people would prefer to contribute only when all reciprocal types meet each other, while in the stranger society reciprocal types may be willing to contribute regardless of the true types of other players.

In Section 4, we revisit the prisoners' dilemma and shed light on cooperation under incomplete information. Rabin (1993) shows that with reciprocal concerns the socially optimal mutual cooperation is achievable when the sensitivity parameters of both players reach certain thresholds. Similarly, the equilibrium strategies in the incomplete information case take the form of threshold strategies. But the thresholds are generally different from those in the complete information case, depending on distributions of players' types. Also, sometimes a small amount of uncertainty in reciprocity motivations can completely eliminate cooperation (Section 4.2.2). Another interesting finding is that when *the advantage of defection is not too large*, the mutual cooperation rate in the acquaintance society is large than that in the stranger society, regardless of type distributions (Section 4.2.1). That means information about people's reciprocal motivation can overall improve mutual cooperation rate. However, when *the advantage of defection is large enough*, information may hinder social cooperation (Section 4.2.3).

In our reciprocity model, the game form of interest is a multi-stage game with observed actions and independent types. As there is incomplete information about types and belief updating at each stage, we need to determine the belief system that is associated with the extensive game form. When the information set is reached with positive probability under a strategy profile, the belief is updated by Bayes rule. The challenge is to determine the beliefs at the histories that occur with zero probability under equilibrium strategy profiles. To address this problem, we inherit the restrictions on the belief system from Kreps and Wilson (1982), that is, we ask the equilibrium assessment to be *consistent*. Then, we adapt the Sequential Reciprocity Equilibrium (SRE), defined in an extensive game form with complete information (DK), to our setup with incomplete information. Note that the SRE is consistent with the *sequential equilibrium* proposed by Battigalli and Dufwenberg (2009), which is a generalization of Kreps and Wilson (1982)'s equilibrium notion.

Sebald (2010) makes the first attempt to extend DK by incorporating chance moves. He stresses that people's procedural concern about whether outcomes are generated through intentional choices or random-

ized processes can greatly affect how they attribute responsibility and how they perceive kindness. He shows that players may use randomized procedures to avoid being held responsible for realized outcomes and thus mitigate others' reciprocal motivations. Unlike this paper, the kind of chance move he focuses on is randomization options players can choose, such as flipping a coin, and he maintains the assumption of complete information about sensitivity parameters. Bierbrauer and Netzer (2016) also analyze uncertainty in reciprocity models in the context of mechanism design. They examine the extent to which implementable social choice functions are robust to the presence of psychological motivations. They provide a condition for the implementability of social choice function when the sensitivity parameters are unknown. In the construction of certain mechanisms, they exploit the feature of reciprocity models that agents' reciprocal incentives are effectively influenced by feasible alternatives.

Finally, our paper sheds light on the problem of public goods provision with reciprocity. Kozlovskaya and Nicolo (2019) find that the *pivotal* mechanism is not strategy-proof in the public goods game when the agents are reciprocal. They demonstrate that once the mechanism is implemented sequentially, the incentive compatibility can be restored. In Section 3.5, we study a public goods game with reciprocal agents where the government imposes taxes on people (the minimum contribution level). Surprisingly, we find that the tax may weaken agents' reciprocal motivation and thus decrease the total contribution level.

The outline of our paper is the following. In Section 2, we describe the reciprocity model with incomplete information in extensive-form games. In Section 3, we apply the theory to a collection of well-known games, and investigate how uncertainty changes the results from reciprocity models with complete information. In Section 4, we study the prisoners' dilemma and illustrate the implications of uncertainty on reciprocal behavior by comparing the equilibrium outcomes between the acquaintance and stranger societies. In Section 5, we conclude.

2. The Model

Our model is composed of three components: the basic game form, the reciprocal payoffs, and incomplete information about sensitivity parameters. The basic game form corresponds to a multi-stage game where players take actions simultaneously at each stage. It consists of the players, their actions, and their material payoffs. The reciprocal payoff of each player depends on his evaluation of other players' intentions.

If the other players are perceived to be kind, he has the incentive to return the favor. But if he feels being mistreated, he would like to damage the unkind players' well-being. As the game proceeds, they will change their understanding of others' intentions as more information is revealed. Finally, the sensitivity parameter is the weight a player puts on his reciprocal payoff. It is assumed to be a random variable that is drawn from a publicly known distribution and privately informed to each player. Below, we will first lay out the basic game form under complete information, and then introduce the uncertainty about sensitivity parameters and the reciprocal payoffs. In the end, we define the equilibrium notion.

2.1. Basic game form under complete information

Suppose there are I ($I \geq 2$) players. The game proceeds for T stages. At each stage, players observe the moves of previous stages and take actions simultaneously. Let h^0 denote the initial history node. At stage 1, each player i has a set of possible actions $D_i(h^0)$. Let h^t denote the history of actions up until stage t, and let $D_i(h^t)$ denote the set of player i's possible actions at h^t. If $D_i(h^t)$ is a singleton, it means that player i is not active at stage $(t+1)$. If at stage $(t+1)$ players take actions $a^{t+1} \in \prod_{i=1}^{I} D_i(h^t)$, then we write $h^{t+1} = (h^t, a^{t+1})$. All the action sets $D_i(h^t)$ are assumed finite. The set of all the histories up to period t is denoted by H^t.

When there is no uncertainty about types, each player's behavioral strategy s_i is a sequence of maps from H^{t-1} to $\Delta(D_i(h^{t-1}))$, where $s_i(h^{t-1})$ is a probability distribution over $D_i(h^{t-1})$. The set of player i's behavioral strategies under complete information is denoted by S_i. We denote by $s = (s_i)_{i=1}^{I}$ the strategy profile of all players and by $S = \prod_{i=1}^{I} S_i$ the set of such strategy profiles. The material payoff function of player i is a mapping $\bar{\pi}_i : S \longrightarrow \mathbb{R}$.

2.2. Uncertainty about sensitivity parameters

In our model, players' psychological motivations (sensitivity parameters) are random variables. We denote by θ_{ij} the sensitivity parameter of player i with respect to player j, which represents the weight of his psychological payoff in player i's utility function. The vector $\theta_i = (\theta_{i1}, \ldots, \theta_{i,i-1}, \theta_{i,i+1}, \ldots, \theta_{iI})$ is the type of player i, randomly drawn from a finite set Θ_i.[3] The product of all type spaces is $\Theta = \prod_{i=1}^{I} \Theta_i$.

[3]The purpose of studying a finite game (finite type spaces and action sets) in the general framework is to guarantee the existence of an equilibrium. In several applications in the following sections where an equilibrium exists, we also allow for continuous type or

At the beginning of the game, nature randomly selects a type θ_i for each player i from the type space Θ_i according to a prior distribution $\mu_i^0 \in \Delta(\Theta_i)$. Each player is privately informed of his own type and has common knowledge of the prior distribution $\mu^0 = \prod_{i=1}^I \mu_i^0$, where the type distributions are independent.

After introducing uncertainty, the domain of player i's behavioral strategies becomes the product of the type and history spaces. Player i's behavioral strategy a_i is a sequence of maps from $\Theta_i \times H^{t-1}$ to $\Delta(D_i(h^{t-1}))$. The set of player i's behavioral strategies is A_i, and the set of the profiles of all players' behavioral strategies is $A = \prod_{i=1}^I A_i$. Since our type is only related to psychological payoffs, the feasible set of actions $D_i(h^{t-1})$ is independent of types. Given $\theta_i \in \Theta_i$ and $a_i \in A_i$, the type θ_i of player i knows that the behavioral strategy he is actually playing coincides with the behavioral strategy under complete information $a_i(\theta_i) \in S_i$ such that $a_i(\theta_i)(h^t) = a_i(\theta_i, h^t)$ for all h^t. Sometimes we also write $a_i(\theta_i)$ as $s_i^{\theta_i}$, and we may suppress dependence on types and write s_i when there is no ambiguity.

Players' (expected) *material* payoffs are written as functions $\pi_i : A \times \Delta(\Theta) \longrightarrow \mathbb{R}$. When players play strategies $a \in A$ and hold a belief $\gamma \in \Delta(\Theta)$, player i's expected material payoff is denoted by $\pi_i(a, \gamma)$. Later we abuse notation by using $\pi_i(s_i^{\theta_i}, a_{-i}, \mu_{-i})$ to denote i's expected payoff when he plays a strategy $s_i^{\theta_i}$ after he knows his type θ_i, given others' strategies a_{-i} and the belief about others' types μ_{-i}.

At each stage of the game, the players will update beliefs about types based on the observed history and initial strategy profile. According to Fudenberg and Tirole (1991), in such a multi-stage game with observed actions, we can assume that they have common knowledge about the beliefs about types formed at each information set, which results in a belief system $\mu : H \longrightarrow \Delta(\Theta)$. In equilibrium, the beliefs in μ are derived from Bayes rule whenever possible and the assessment (a, μ) should satisfy the property of *consistency* as defined in Kreps and Wilson (1982).

Definition 1. *An assessment (a, μ) is* consistent *if there is a sequence of completely mixed behavioral strategy profiles $a^n \longrightarrow a$ such that $\mu^n \longrightarrow \mu$, where μ^n is the belief system induced by a^n for all n.*

action spaces. These applications include the prisoners' dilemma, the battle of the sexes, and the ultimatum game.

2.3. Reciprocity payoffs

The kindness of player i to other players is reflected by the consequences of his choices on others' well being, and therefore depends on the behavioral strategy $s_i^{\theta_i}$ of the realized type θ_i of player i, and on his belief about other players' strategies, $(b_{ij})_{j\neq i} \in \prod_{j\neq i} A_j$ (first-order beliefs). Conversely, the kindness of player j to i, from i's point of view, depends on i's belief about player j's strategy $b_{ij} \in A_j$ and i's belief about j's belief about all other players' strategies $(c_{ijk})_{k\neq j} \in \prod_{k\neq j} A_k$ (second-order beliefs). We denote the set of players' first-order beliefs by B_{ij} $(= A_j)$, for all i and $j \neq i$ and denote the set of players' second-order beliefs by C_{ijk} $(= B_{jk})$, for all i, $j \neq i$, and $k \neq j$.

As game unravels, each player would reevaluate his and other players' kindness based on new information. At each history h, he would hold other players responsible for all the decisions they have made and treat them as if they had made those decisions deliberately with probability 1. Following DK, we define the updated belief about player i's strategy at history h to be $a_{i,h}$, which has the property that it equals a_i at all histories except those that define h. That is, in all the predecessors of h, $a_{i,h}$ assigns probability 1 to the actions that leads to h for all types. Similarly, players also revise their first- and second-order beliefs from b_{ij} and c_{ijk} to $b_{ij,h}$ and $c_{ijk,h}$ at history h. If it is a strategy for type θ_i, $s_i \in S_i$, it is revised at each history h in the analogous manner and denoted by $s_{i,h}$.

In DK, the interpretation of this *revision of the beliefs about strategies* is that the randomization prescribed by the strategy at some history represents the frequencies with which pure actions are taken in a "population." Since information is usually insufficient to identify the individual, players still need to update beliefs about types for the purpose of assessing kindness in the continuation game. It is worth noting that the belief updating of types is based on the *initial* strategies but not the *revised* strategies. Otherwise, there will be no change in beliefs about types at any information set regardless of whether it is on or off the equilibrium path.

Since the kindness of each player i is measured by the intended consequences of what he does relative to what he could have done, the set of the alternative options from which he can choose plays an important role in the assessment of his kindness. In particular, we focus on reasonable strategies they expect others to play, which are defined as *efficient strategies*. We will first define an *efficient strategy* for a single type of a player in the same way as in complete information games (Dufwenberg and Kirchsteiger, 2019). The idea

is that, for any type of a player, an efficient strategy should not be outperformed by another strategy in all histories for any strategies played by other players.

The set of efficient strategies for player i under complete information is defined as below:

$$E_i = \{s_i \in S_i |\ \nexists s_i' \in S_i \text{ such that for all } h \in H \text{ and } (s_j)_{j\neq i} \in \prod_{j\neq i} S_j,$$

$$\bar{\pi}_i(s_{i,h}', (s_{j,h})_{j\neq i}) \geq \bar{\pi}_i(s_{i,h}, (s_{j,h})_{j\neq i}), \text{ with at least one strict inequality}\}$$

Then, we restrict efficient strategies to comply with efficiency for each type. Thus, the set of efficient strategies is denoted by $\tilde{E}_i = \{a_i \in A_i | a_i(\theta_i) \in E_i \text{ for each } \theta_i \in \Theta_i\}$.

We now provide the definition of kindness under private information. For each type, player i's kindness is measured, analogously to the case of complete information, by the effect of her action on the (expected) material payoff of player j. Since her action depends on her type, her kindness is also type-dependent.

In any continuation game, each player i updates his belief about other players' types to $\gamma_{-i} \in \prod_{j\neq i} \Delta(\Theta_j)$ and his belief about other players' strategies to $(b_{ij})_{j\neq i} \in \prod_{j\neq i} B_{ij}$. Under these beliefs, he knows the range of possible expected payoffs to player j ($j \neq i$) as he varies his strategy s_i. That is, $\{\pi_j(s_i, (b_{ij})_{j\neq i}, \gamma_{-i}) | s_i \in S_i\}$. In the view of player i, the equitable payoff for j is the *average* expected payoff to j caused by player i's choice conditional on her type θ_i.

$$\pi_{ij}^e((b_{ij})_{j\neq i}, \gamma_{-i}) = \frac{1}{2}\Big[\max\big\{\pi_j(s_i, (b_{ij})_{j\neq i}, \gamma_{-i}) | s_i \in S_i\big\} + \min\big\{\pi_j(s_i, (b_{ij})_{j\neq i}, \gamma_{-i}) | s_i \in E_i\big\}\Big]$$

Note that when calculating the minimum payoff we exclude non-efficient strategies from consideration. A simple example that provides an explanation for this setup is included in Figure 3 of DK.

Then, player i's kindness to j is measured by the expected payoff to j from i's point of view as compared to the equitable payoff. Formally, the kindness of i to j at h is a function $\kappa_{ij} : S_i \times \prod_{j\neq i} B_{ij} \times \Delta(\Theta) \to \mathbb{R}$ defined by

$$\begin{aligned}&\kappa_{ij}(s_{i,h}, (b_{ij,h})_{j\neq i}, \mu_{-i}(h)) \\ =&\pi_j(s_{i,h}, (b_{ij,h})_{j\neq i}, \mu_{-i}(h)) - \pi_{ij}^e((b_{ij,h})_{j\neq i}, \mu_{-i}(h))\end{aligned}$$

On the other hand, there are two standards for player i to evaluate the equitable payoff she deserves from j based on whether i cares about the *ex post* consequence of j's behavior for her realized type or the overall consequence for her *ex ante* payoff. We adopt the second interpretation for the reason that player j does not have knowledge about i's type, so i should not hold j accountable for the effect of j's strategy solely on the payoff of the realized type of i. In Appendix A.1, we have a detailed discussion on disadvantages of the alternative definition.

Given that player i cares about the overall effect of j's behavior, the equitable payoff to i depends on j's behavioral strategy a_j. Suppose from player i's point of view, j thinks that other players are taking strategies $(c_{ijk})_{k\neq j}$. Then, even though j does not know i's realized type and chosen action, he knows that his behavioral strategy will effectively determine i's expected payoff. As below, the equitable payoff to player i from j is defined as the average of the maximum and minimum of the expected payoffs that player i believes player j could have given to him given their belief about type distribution $\gamma \in \Delta(\Theta)$.

$$\begin{aligned}\pi^e_{iji}((c_{ijk})_{k\neq j},\gamma) = \frac{1}{2}\Big[&\max\left\{\pi_i(a_j,(c_{ijk})_{k\neq j},\gamma)|a_j\in A_j\right\}\\ &+\min\left\{\pi_i(a_j,(c_{ijk})_{k\neq j},\gamma)|a_j\in \tilde{E}_j\right\}\Big]\end{aligned}$$

Next, player i would perceive the kindness of j to himself as his expected payoff resulting from j's play relative to his equitable payoff. Formally, his *perceived kindness* is a function $\lambda_{iji} : B_{ij}\times\prod_{k\neq j}C_{ijk}\times\Delta(\Theta)\to\mathbb{R}$. Specifically,

$$\begin{aligned}&\lambda_{iji}(b_{ij,h},(c_{ijk,h})_{k\neq j},\mu(h))\\ =&\pi_i(b_{ij,h},(c_{ijk,h})_{k\neq j},\mu(h)) - \pi^e_{iji}((c_{ijk,h})_{k\neq j},\mu(h))\end{aligned}$$

Our definitions of kindness and perceived kindness are in the spirit of Sebald (2010), in that a player's kindness should be determined by the player's intention instead of based on chance moves. Information asymmetry explains why there is asymmetry between the definitions of kindness and perceived kindness. Each player knows his own type, so he should hold himself responsible for the consequences of the actions associated with his type. But other players do not know which type he is, so he does not hold them responsible for the consequences solely on the payoff of the realized type he knows.

To capture player i's motivation *to be kind to the kind and unkind to the unkind*, we write down player i's reciprocal payoff toward j as the product of the kindness and unkindness terms, multiplied by a sensitivity parameter. The (expected) utility of player i of type θ_i at h is a function $U_i^{\theta_i} : S_i \times \prod_{j\neq i}\left(B_{ij} \times \prod_{k\neq j} C_{ijk}\right) \times \Delta(\Theta) \to \mathbb{R}$, which can be separated into material and psychological payoffs. Specifically,

$$U_i^{\theta_i}(s_{i,h}, (b_{ij,h}, (c_{ijk,h})_{k\neq j})_{j\neq i}, \mu(h)) =$$

$$\pi_i(s_{i,h}, (b_{ij,h})_{j\neq i}), \mu(h)) + \sum_{j\neq i} \theta_{ij} \cdot \kappa_{ij}\left(s_{i,h}, (b_{ij,h})_{j\neq i}, \mu_{-i}(h)\right)\lambda_{iji}(b_{ij,h}, (c_{ijk,h})_{k\neq j}, \mu(h))\Big)$$

2.4. Equilibrium

In the equilibrium analysis, we treat each player as being a rational "agent" at different histories. Each agent (i,h), together with his utility, is identified with the corresponding player and the history at which the player makes a move. The equilibrium assessment requires that players update beliefs in the above way and that each agent maximize his "local" utility.

Before we give the definition of equilibrium, we introduce another piece of notation $S_i(\theta_i, h, a) \subseteq S_i$, which contains strategies that prescribe the same actions for type θ_i as $a_i(\theta_i, h')$, at every history h' except h.

Finally, we define the equilibrium notion as below.

Definition 2. (a^*, μ^*) *is a Sequential Reciprocity Equilibrium (SRE) if:*

(1) At each history h, for each player i and each θ_i*, the following conditions are satisfied:*

(1.1) $a^*_{i,h}(\theta_i) \in \arg\max_{s_i \in S_i(\theta_i,h,a^*)} U_i^{\theta_i}(s_i, (b_{ij,h}, (c_{ijk,h})_{k\neq j})_{j\neq i}, \mu^*(h))$*;*

(1.2) $b_{ij} = a^*_j$*, for all* $j \neq i$*;*

(1.3) $c_{ijk} = a^*_k$*, for all* $j \neq i$*,* $k \neq j$*.*

(2) (a^*, μ^*) *is* consistent *according to Definition 1.*

According to Definition 2, condition (1.1) states that at each history h, player i maximizes his utility given his updated beliefs about types and the equilibrium strategy profile of all players. Conditions (1.2) and (1.3) state that all players update their beliefs correctly, and therefore the first- and second-order beliefs coincide with the equilibrium strategy profile. Condition (2) requires that the assessment satisfies *consistency* proposed by Kreps and Wilson (1982), which means that the belief updating obeys Bayes' rule at each

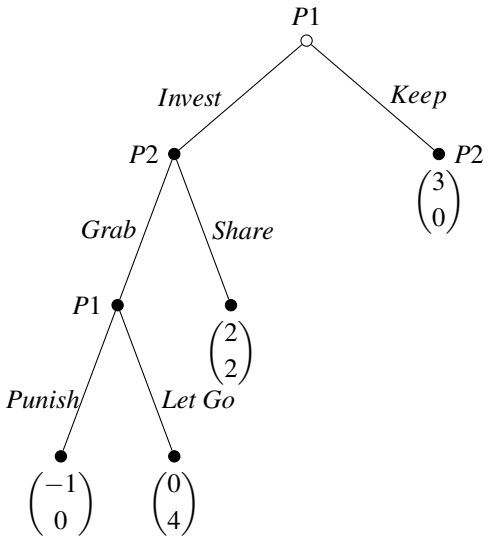


Figure 1: Investment Game with Punishment

information set that is reached with positive probability and that in "zero-probability" events players would hold beliefs that can be approximated by a sequence of beliefs induced by completely mixed strategies.

The SRE is a special case of the equilibrium notion provided in Battigalli and Dufwenberg (2009), who generalize the *sequential equilibrium* by Kreps and Wilson (1982) to psychological games. Therefore, we can rely on their existence proof to demonstrate the existence of an SRE. We also provide an existence proof of SRE for the reciprocity model in Appendix A.2.

Theorem 1. *In any psychological game with reciprocity motivations, an SRE exists.*

3. Applications

In this section, we apply the theoretical model to a series of games and illustrate the implications of uncertainty for reciprocal behavior. We also compare the equilibrium outcomes in *acquaintance* and *stranger* societies, where players start the game with and without the knowledge of each other's types, respectively.

3.1. Investment Game with Punishment

In this subsection, we study an investment game with punishment. There are two players, an investor (P1) and an entrepreneur (P2). The investor chooses whether or not to invest in a project managed by the entrepreneur. If she decides to invest, there could be two possibilities. First, the entrepreneur may put in efforts and share the gain with her. Second, the entrepreneur may claim all of the benefit and leave the investor nothing; In this case, the investor can further choose to engage in costly punishment. The game is illustrated in Figure 1.

The unique Subgame Perfect Equilibrium in the standard game is (*Keep*, *Let Go*; *Grab*). However, in reciprocity games, (*Invest*, *Share*) could possibly be supported as an equilibrium outcome. We are interested in the conditions under which (*Invest*, *Share*) can be supported as an equilibrium outcome.

Suppose P2 (he) is selfish but P1 (she) could be either reciprocal or selfish. Let θ denote P1's sensitivity parameter and let p denote the probability that P1 is reciprocal. The only reason for which P2 might choose *Share* is that he has concerns that P1 would punish him for choosing *Grab*.

Note that P1's investment decision can credibly signal her type. In the first stage, the selfish type of P1 will certainly choose *Keep*, which gives her the highest payoff that can be achieved in the game. That means whenever P1 chooses *Invest*, P2 would understand that it is the reciprocal type of P1 that makes this choice. This changes P2's anticipation of P1's punishment behavior — P2 knows that only the reciprocal type of P1 could possibly punish him if he decides to *Grab*.

However, P1 would choose to punish P2 only if she has strong enough psychological motivations. Let us calculate the lower bound for her sensitivity parameter. At history (*Invest*, *Grab*), players would revise their beliefs about the strategy profile to that P1 takes *Invest* with probability 1 and P2 takes *Grab* with probability 1. Under this belief and the belief that P1 would choose *Punish*, P2 is unkind to P1 by choosing *Grab*, and the unkindness is equal to $-\frac{3}{2}$. As for P1, she is kind to P2 if she chooses *Let Go* ($\kappa_{12} = 2$), and unkind if she chooses *Punish* ($\kappa_{12} = -2$). Therefore, the condition for P1 to choose *Punish* is:

$$-1 + 3\theta \geq -3\theta \Rightarrow \theta \geq \frac{1}{6} \tag{1}$$

At history (*Invest*), P2 would choose *Share* only if he thinks that P1 would punish him with probability

more than $\frac{1}{2}$. Since in equilibrium only the reciprocal type of P1 can choose *Invest*, P2's choice depends on the strategy of P1's reciprocal type. When $\theta \geq \frac{1}{6}$, the reciprocal type of P1 would punish P2 so that P2 would share conditional on (*Invest*).

At the first stage, P1 would choose *Invest* only when she believes P2 would kindly share with her. Under this belief, P1's kindness to P2 by taking *Invest* equals 1, and unkindness by taking *Keep* equals -1. Note that when P1 assesses the consequence of P2's action on her expected payoff, she takes into account the fact that the selfish type of herself always chooses *Keep*. If the reciprocal type of P1 chooses *Invest* and P2 chooses *Share*, P1's expected payoff equals $3(1-p)+2p$; while if P2 chooses *Grab*, P1's expected payoff equals $3(1-p)-p$. Therefore, the kindness of P2 by taking *Share* is $\frac{3}{2}p$. Thus, the condition for the reciprocal type of P1 to choose *Invest* is:

$$2+\theta\cdot\frac{3}{2}p \geq 3-\theta\cdot\frac{3}{2}p \Rightarrow \theta \geq \frac{1}{3p} \tag{2}$$

Since $p \in (0,1]$, condition (2) is stronger than condition (1), and we have the following result.

Observation 1. *When $\theta \geq \frac{1}{3p}$, there exists a reciprocity equilibrium where the reciprocal type of P1 takes (*Invest, Punish*), the selfish type of P1 takes* Keep*, and P2 takes* Share*.*

We can also make a comparison with the reciprocity model without uncertainty. When $p=1$, the game lies in the realm of DK. Based on similar analyses, we find that when $\theta \geq \frac{1}{3}$, (*Invest*, *Punish*; *Share*) is a reciprocity equilibrium. From condition (2), we can see that the threshold is generally higher in the incomplete information case, which means that P1's psychological motivations weaken with the increase in the probability that P1 is selfish. Because a lower value of p means that it is more likely that P1 would choose *Keep* and thus P2's reaction is relatively less important for her. After introducing uncertainty about P1, the requirement for her to make the decision that could lead to the socially optimal outcome becomes higher.

3.2. Ultimatum Game

In an ultimatum game, there is a proposer (P) and a responder (R). We will use male pronoun (he) for P and female pronoun (she) for R. P offers a split of a unit pie into $(1-x,x)$, $x \geq 0$, and R decides to *accept* or

reject. If R accepts the offer, she will receive a payment x, and P will receive $1-x$. Otherwise, they get zero payoffs.

P's pure strategy is a number $x \in [0,1]$. R's pure strategy is a mapping $s : [0,1] \to \{A,R\}$. For simplicity, we suppose that R plays a threshold strategy with a cutoff level $\bar{s}$. Under this strategy, she will accept if $x \geq \bar{s}$ and reject otherwise. We focus on the situations where R accepts the offer on the equilibrium path. Also, it turns out that there is a continuum of equilibria, while we only study P's favorite equilibrium when analyzing comparative statics.[4]

Suppose R's sensitivity parameter is θ_R. When it is public information, P's favorite equilibrium is featured by an offer $\bar{s}(\theta_R)$ to R. The following result shows the desirable properties of this function.

Proposition 1. *In P's favorite equilibrium, he would offer R an amount equal to*

$$\bar{s}(\theta_R) = \frac{1}{4}\left[(3+\frac{2}{\theta_R}) - \sqrt{(3+\frac{2}{\theta_R})^2 - 8}\right] \tag{3}$$

$\bar{s}(\theta_R)$ is monotonically increasing in θ_R. Specifically, when $\theta_R \longrightarrow 0$, $\bar{s}(\theta_R) \longrightarrow 0$; when $\theta_R \longrightarrow \infty$, $\bar{s}(\theta_R) \longrightarrow \frac{1}{2}$.

Now we turn to an incomplete information game with binary types for R, a selfish type with $\theta_R = 0$ and a reciprocal type with $\theta_R > 0$. Let the prior probability of the reciprocal type be $\mu \in (0,1)$. In contrast, P is assumed to be selfish. In equilibrium, the selfish type of R plays a strategy that prescribes acceptance in response to any offer. We only need to discuss the choice of the unselfish type of R. Let $\bar{s}(\theta_R)$ be the cutoff level for the reciprocal type as in the complete information game with $\theta_R > 0$.

It turns out that there are only two possible offers made by P in P's favorite equilibrium, $x = 0$ or $x = \bar{s}(\theta_R)$. If $x = 0$, the selfish type would accept and the unselfish type would reject, so that P's expected payoff is $(1-\mu)$. If $x = \bar{s}(\theta_R)$, both types would accept and P's expected payoff is $1 - \bar{s}(\theta_R)$.

Is it possible that P offers something in between 0 and $\bar{s}(\theta_R)$? No, this is a kind of offer that the unselfish type would reject and the selfish type would accept. It is obviously more profitable for P to deviate to $x = 0$.

[4]The reason that there could be multiple cutoff levels for equilibrium lies in that the kindness of P depends on his expectation of R's response. If he gives a positive offer which he expects R to accept, it would be nice of him; but for the same offer, if he expects it to be rejected, then this move should be regarded as unkind. Therefore, it is possible that a positive offer is chosen by P and accepted by R in one equilibrium but rejected in another equilibrium.

Proposition 2. *P's equilibrium choice is either $x = 0$ or $\bar{s}(\theta_R)$, which depends on the comparison between $(1-\mu)$ and $1-\bar{s}(\theta_R)$. Because $\bar{s}(\theta_R) < \frac{1}{2}$ for all $\theta_R \in [0,+\infty)$, when $\mu \geq \frac{1}{2}$, P will offer $\bar{s}(\theta_R)$.*

How should we understand that when μ is large enough, P would give an offer $\bar{s}(\theta_R)$ as high as that towards a reciprocal R under complete information? Originally, it is negative reciprocity that drives a positive offer and it seems that the presence of the selfish type may weaken the strength of reciprocal motivations and result in a lower offer. But what matters is that the reciprocal type of R would not compromise on the amount she should receive for her to accept it. That is because no matter whether R is reciprocal or not, P can always give the whole pie to her or make her earn nothing. So the equitable payoff to R is left unchanged for both types. Then, under the expectation that P would like both types of R to accept the offer, P's intention of making any offer is perceived the same as that under complete information. Thus, the uncertainty of R's type does not affect R's perception of P's kindness, and the reciprocal type of R would reciprocate as if P has already known her type. On the other hand, the selfish type of R can "free ride" on the reciprocal type's high requirement and receive this positive offer.

3.3. Monopoly Pricing

One context in which reciprocity has been discussed is *monopoly pricing*. Rabin (1993) has shown that when consumers are motivated by reciprocity,[5] they might refuse to buy from the monopolist if the price is higher than what they deem fair. The profit-maximizing monopolist would, therefore, set the price lower than what models with self-interest preferences predict. We will apply our framework to explore how monopoly pricing would change when the firm is unsure of the consumers' reciprocal motivations.

Consider a profit-maximizing monopolist (M) who produces a good which costs c per unit, and a typical consumer (C) whose valuation for the good is $v > c$. Without loss of generality, let $v - c = 1$. M can choose the price $p \in [c, v]$. C may *buy* or *refuse* to buy. If C buys, M gets $p - c$ and C gets $v - p$. Otherwise, they both get 0. There are two types of C: $\theta_C \in \{0, \theta\}$, where $\theta > 0$ and $Prob(\theta_C = 0) = \eta \in (0,1)$. We shall call the type $\theta_C = 0$ *the selfish type* and call $\theta_C = \theta$ *the reciprocal type*. Everything except C's type is common knowledge. Suppose that the selfish type will buy at $p = v$ to break the tie. Then, we will focus on the cutoff

[5] Rabin (1993) calls it fairness.

strategies for the reciprocal type. That is, the reciprocal type of C will choose a reservation price $r \in [c, v]$ such that if $p \leq r$, he will buy and otherwise he will not. Thus, an equilibrium is described by a pair (p, r).[6]

It turns out that there is a continuum of equilibria in which the reciprocal type's reservation price ranges from $\underline{p} := v - \frac{(2\eta+1+2/\theta)+\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta}$ to $\bar{p} := v - \frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4}$. It can be checked that $c < \underline{p} < \bar{p} < v$.

Observation 2. *In any equilibrium, the reservation price r for the reciprocal type of C lies in* $[\underline{p}, \bar{p}]$.

Given any reservation price r, M can either decide to price it at r and have both types buy it, or price at v to maximally take advantage of the selfish type and let the reciprocal type *refuse*. If η is sufficiently small, M will charge the reservation price r. If the probability of the selfish type η is sufficiently large, M will charge $p = v$.

Therefore, if the probability of the selfish type is large, Rabin's (1993) insight that M will offer a lower price than v does not hold. When $\eta > \frac{1}{2}$, it may be beneficial for M to set $p = v$ to exploit the selfish type to the full extent. It is then natural to ask under what condition will M set price below v. A sufficient condition for $p < v$ to occur is that C is more likely to be *reciprocal* than *selfish*.

Observation 3. *If the probability of the selfish type is* $\eta \leq \frac{1}{2}$, *there is an equilibrium where both types buy at a price* $p < v$.

Under this sufficient condition ($\eta \leq \frac{1}{2}$), we shall compare the highest-reservation-price equilibria and the lowest-reservation-price equilibria between the stranger and the acquaintance societies. In the acquaintance society, the highest reservation price is $r = \bar{p}$ and M will set the price $p = \bar{p}$. And the lowest reservation price is $r = \underline{p}$ and M will set the price $p = \underline{p}$. In the stranger society, the highest reservation price is the same as that in the acquaintance society.

Proposition 3. *If* $\eta \leq \frac{1}{2}$,

(i) When the highest-reservation-price equilibria are compared, M's equilibrium price in the stranger society is identical to the price for the reciprocal consumer in the acquaintance society.

[6] We suppress the beliefs since they must coincide with the strategies in equilibrium.

(ii) When the lowest-reservation-price equilibria are compared, M's equilibrium price in the stranger society is strictly higher than the price for the reciprocal consumer in the acquaintance society.

One takeaway from (i) is that if more than a half of the population is reciprocal, the selfish individual will benefit in the stranger society, as they will face a price lower than that in the acquaintance society. The reciprocal individuals are not worse off as they face the same price as in the acquaintance society. In short, selfish consumers are better off in the stranger society since they are treated like the reciprocal types due to unobservability of types. This may not necessarily hold if $\eta > \frac{1}{2}$. Recall that when η is large, M will charge $p = v$ so that only the selfish type will buy at the highest price possible. In this case, the reciprocal type will face a lower price in the acquaintance society.

The intuition for (ii) is as follows. It claims that C may agree to accept a higher price when C and M are strangers than when they are acquaintances. In other words, there could be a price $p < v$ that would be rejected in the acquaintance society but would be accepted in the stranger society. The reason is that such a price is *not* low enough to be deemed kind by the reciprocal type of C so that he would prefer to reject it. In the acquaintance society, by offering p, M should know for sure that it will be rejected, so he is unkind to the reciprocal type of C. In the stranger society, however, M might have expected that the price could possibly be accepted by the selfish type of C, which reduces M's unkindness. Thus, the reciprocal type of C may not have strong enough reciprocal motivations to reject the price.

3.4. Sequential Prisoners' Dilemma

We revisit the sequential prisoners' dilemma studied by DK (Figure 2). In this sequential prisoners' dilemma, the second player (she) can condition his decision on the first player (he)'s choice. Suppose player i's sensitivity parameter is a random variable that takes on values either 0 or θ_i (> 0). The prior probability of θ_i is p_i. We examine under what conditions the cooperative equilibrium, where the unselfish types of both players 100% choose C, can be supported.

In the second round, conditional on the decision node following D, P2 revises her belief about P1's strategy such that she would treat P1 as if he was playing a strategy that assigns probability 1 to D. Based on this belief, she would view P1 as unkind, and choose d irrespective of her type. On the contrary, if the decision node following C is reached, the reciprocal type of P2 may want to choose c when she is strongly

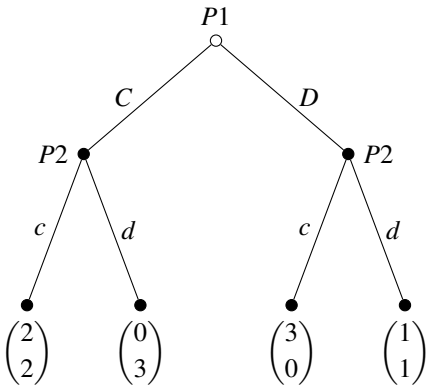


Figure 2: Sequential Prisoners' Dilemma

motivated by reciprocity.

Observation 4. *Given that P1 has chosen C, it is optimal for the reciprocal type of P2 to choose c if* $\theta_2 \geq \frac{1}{2-p_2}$.

Referring to DK, in complete information games, the threshold for P2 to cooperate conditional on P1's cooperation is $\theta_2 \geq 1 (> \frac{1}{2-p_2})$. It means that P2 has a stronger incentive to cooperate when she knows that P1 does not know her type when he chooses *C*. That is because P2 understands that, when P1 chooses *C*, he faces a risk that P2 might be selfish and would choose *d* in response. In this case, she is more grateful for P1's generosity when he chooses *C*.

Then, we discuss the conditions under which the reciprocal types of both players can choose cooperation in equilibrium.

Observation 5. *Case 1: When* $p_2 \geq \frac{1}{2}$ *and* $\theta_2 \geq \frac{1}{2-p_2}$*, there is an equilibrium where both types of P1 cooperate and the reciprocal type of P2 cooperates conditional on C.*

Case 2: When both players are more likely to be selfish ($p_1, p_2 < \frac{1}{2}$*), then cooperation will not happen.*

Case 3: When $p_1 \geq \frac{1}{2}$ *and* $p_2 < \frac{1}{2}$*, then there is an equilibrium where the reciprocal type of P1 cooperates, the selfish type of P1 defects, and the reciprocal type of P2 cooperates conditional on C, given that*

both players have strong enough reciprocal motivations. Specifically, it needs to be satisfied that

$$\theta_1 \geq \frac{1-2p_2}{(2p_1-1)(2-p_2)}$$
$$\theta_2 \geq \frac{1}{2-p_2}.$$

The above results show that the prior probabilities of reciprocal types are essential for the cooperation between the two players. First, cooperation is impossible if both players are more likely to be selfish. Second, notice that the thresholds of θ_1 and θ_2 are decreasing in p_1 and p_2, respectively. It means that the more possible they are believed to be reciprocal, the easier for their sensitivity parameters to reach the requirement for cooperation.

Observation 6. *If $p_2 > \frac{1}{2}$, the probability of reaching mutual cooperation is weakly higher in the stranger society than in the acquaintance society.*

This result derives directly from Observation 5. When P2 is more likely to be reciprocal, her incentive to cooperate conditional on P1's cooperating is strengthened in the stranger society. That means, if the reciprocal type of P2 does not want to cooperate in the stranger society, she would not like to cooperate in the acquaintance society either. Furthermore, since $p_2 > \frac{1}{2}$, once P2's reciprocal type wants to cooperate, it is optimal for P1 to cooperate regardless of his type, because his material expected payoff from playing C is higher. In addition, P2 is overall kind to P1, so he wants to return the favor by taking C.

3.5. Public Goods Game

We revisit the public goods game with reciprocity concern studied by Dufwenberg, Gachter and Hennig-Schmidt (2011) and investigate what impact uncertainty has on the level of public goods provision. In addition, we look at how a mandatory minimum contribution affects the total contribution level when the players have reciprocal motivations. Several experimental studies have explored the effect of a mandatory minimum contribution in the public goods game.[7] In particular, Martinsson, Medhin and Persson (2019) has reported

[7] Andreoni (1993) and Gronberg, Luccasen III, Turocy and Van Huyck (2012) experimentally study the effect of a tax on the contribution level. Their study uses *non-linear* production functions that differ from the standard public goods game. The non-linear production function makes some positive contribution the dominant strategy for selfish players. They find that taxes increase the total contribution level. In contrast, Kocher, Martinsson, Persson and Wang (2016) report that, using the linear production function, tax does not decrease the total contribution level.

that introducing a mandatory minimum contribution can surprisingly decrease the total contribution level. However, the theoretical relationship between a mandatory minimum contribution and reciprocal agents' incentives for contribution has not been explored. Our theory provides an explanation for this phenomenon.

Suppose there are three players, each of whom is endowed with 20 tokens. Each player i can contribute $s_i \in [x, 20]$ to producing the public good G with the production function $G = \frac{3}{2}(s_1 + s_2 + s_3)$. The value $x \geq 0$ is the minimum required contribution (tax). All the public goods produced are equally shared by the players so that each receives $\frac{1}{3}G = \frac{1}{2}(s_1 + s_2 + s_3)$. The material payoff to player i equals the remaining tokens, $20 - s_i$, plus public goods he receives, $\frac{1}{3}G$.

Let us introduce uncertainty by assuming that the type of each player θ_i could take on two values, $\underline{\theta}_i$ and $\bar{\theta}_i$ ($\underline{\theta}_i < \bar{\theta}_i$), where the prior probability of $\bar{\theta}_i$ equals $p \in (0, 1]$. To study an interesting case, we assume that $\underline{\theta}_i < \frac{2}{20-x} < \bar{\theta}_i$ for every i, so that high types would like to contribute given that the other two players would make full contributions, but all low types would only contribute the minimum amount possible.[8]

Observation 7. *The incentive constraint for type θ_i of player i to contribute his tokens is*

$$\theta_i \Big[\mathbb{E}(b_{ij}) + \mathbb{E}(b_{ik}) - 20 - x \Big] \geq 2 \tag{4}$$

Eq. (4) shows that in order for a player to contribute, the player needs to be sufficiently sensitive and at the same time, she must believe that other players are contributing more than $20 + x$ in expectation.

Because only high types may make some contribution, a low value of p tends to lower expectation about each other's contribution, which further lowers incentives for contribution. As a result, a low value of p lowers the chance of socially optimal outcome, (i.e., everybody contributes all of their tokens). There could be no contribution and partial contribution in equilibrium, depending on the values of $\bar{\theta}_i$ and p.

Observation 8. *The symmetric equilibria may take two forms:*

1. *If $p \leq \frac{1}{2}$, then no player contributes to the public good in any equilibrium.*

2. *If $p > \frac{1}{2}$, then there is an equilibrium where type $\underline{\theta}_i$ of all players contribute 0, and type $\bar{\theta}_i$ contribute*

[8] If $\bar{\theta}_i < \frac{2}{20-x}$, then there is no equilibrium with a positive contribution level. If $\frac{2}{20-x} < \underline{\theta}_i$, there is an equilibrium where every type of every player fully contributes.

20 tokens when

$$\bar{\theta}_i \geq \frac{2}{20(2p-1)-x}. \tag{5}$$

The first case of the observation states that because the low type never contributes, when the probability of the low type is greater than $\frac{1}{2}$, no other player is considered *kind* on average. Therefore, no one contributes. In the second case, there exists a symmetric equilibrium where all players' high types fully contribute only when the sensitivity parameter of each player satisfies $\theta_i \geq \frac{2}{20(2p-1)-x}$. Note that the threshold is increasing in x; the higher the tax, the higher the threshold for a full contribution. So, increasing x may lead to a significant reduction in the total contribution level. For example, suppose that $p = 1$ and $\bar{\theta}_i \in [\frac{1}{10}, \frac{1}{9})$, for all i. Then, when $x = 0$, according to Observation 8, there is an equilibrium where all players contribute all of their tokens. However, when $x = 1$, it turns out that no one contributes any more than x. This is because as x increases, the contributing behavior becomes less kind for other players, and each player has a weaker motivation to contribute.

Now, we provide a sufficient condition under which removing the tax entirely increases a total contribution level.

Observation 9. *Suppose that the social planner has an option to levy the players a tax of $x > 0$ or $x = 0$. When $20p > x$, and when there exists $i \in \{1,2,3\}$ such that $\bar{\theta}_i \in [\frac{2}{20(2p-1)}, \frac{2}{20(2p-1)-x})$, the expected contribution level is higher under $x = 0$ than under $x > 0$.*

If the high type of player i lies in the interval, $[\frac{2}{20(2p-1)}, \frac{2}{20(2p-1)-x})$, the type $\bar{\theta}_i$ would like to contribute when there is *no* tax, but would not when there is a tax. There is a trade-off between an increase in the contribution of the low type $\underline{\theta}_i$ and a decrease in the contribution of the high type $\bar{\theta}_i$. If $20p > x$, the gain from the high type by lifting the tax outweighs the loss from the low type, so the total contribution level increases. This example has a design implication. If the social planner's goal is to increase the level of public goods, removing the tax may significantly increase the public goods provision, given that the prior probability of highly sensitive types is sufficiently high.

In the following, we compare the stranger and acquaintance societies in the public goods game. If we focus on the equilibrium with the highest contribution levels, then we have the following result.

Observation 10. *If $p > \frac{1}{2}$ and $\bar{\theta}_i > \frac{1}{20(2p-1)-x}$ for all i, then the stranger society has a higher contribution*

level than the acquaintance society in expectation.

In this case, the stranger society outperforms the acquaintance society because in the stranger society, the high type of each player would like to contribute 20 and the low type x, which results in the expected total contribution level as $60p + 3(1-p)x$. In the acquaintance society, however, the high type of each player wants to contribute only when all other two players are of high types. Otherwise, they all contribute the minimum amount x. So the expected total contribution level is $60p^3 + 3(1-p^3)x$, which is less than that in the stranger society.

3.6. battle of the sexes

	Yield	*Assert*
Yield (Y)	$0,0$	$1,3$
Assert (A)	$3,1$	$0,0$

Figure 3: Battle of the Sexes

In this section, we study the battle of the sexes[9] in Figure 3 and discuss the implications of reciprocal uncertainty.[10] In the standard battle of the sexes with selfish preferences, the only equilibrium outcomes (*Y, A*) and (*A, Y*) are Pareto efficient. However, a reciprocity model allows the inefficient outcomes (*Y, Y*) and (*A, A*) to be supported as equilibrium outcomes *under complete information*. We show that this result may not be robust under uncertainty. We will study an example in which a small perturbation of players' type distribution can dramatically change the probability of reaching the outcome (*Y, Y*) in equilibrium. On the other hand, (*A, A*) turns out to be robust in our example. A small perturbation in the players' beliefs about others does not reduce the probability of reaching (*A, A*) much.

First, we study the complete information game as the benchmark. Call the players Man (M) and Woman (W) with the types θ^M and θ^W respectively. The following observation summarizes the equilibrium predictions under complete information.

Observation 11. (i) *(Y, A) and (A, Y) are equilibria for all θ^M and θ^W.*

[9] Since full characterization of equilibria involves lengthy calculations, we will discuss further in Appendix B.6.

[10] Note that while the battle of the sexes is typically presented as an asymmetric game form as the players disagree on their preferred activity, the strategies can be relabeled so it is symmetric as in Figure 3. Yield refers to going to the activity that the co-player prefers (e.g., Man going to Ballet or Woman going to Football) and Assert refers to the action of going to the activity one prefers him/herself.

(ii) *(Y, Y) is an equilibrium if* $\theta^M, \theta^W \geq 6$.

(iii) *(A, A) is an equilibrium if* $\theta^M, \theta^W \geq \frac{2}{9}$.

Observation 11-(i) follows from that in the Pareto efficient outcomes (*Y, A*) and (*A, Y*), players are maximizing their material payoffs and being kind to the kind co-player. For inefficient outcomes (*Y, Y*) and (*A, A*) to be supported in equilibrium, they must be sufficiently reciprocal to have incentives to punish each other for being unkind. Furthermore, (*Y, Y*) requires a stronger reciprocity sensitivity than (*A, A*), as the players forgo more material payoffs to punish the opponents in equilibrium (*Y, Y*), as compared to in equilibrium (*A, A*). In addition, by both playing Y, the players are not as unkind to each other, which weakens the psychological motivations for punishment.

Now, consider the incomplete information game. Suppose that players' reciprocal types, θ^M and θ^W, are random variables distributed according to cumulative distribution functions, F^M and F^W, with support $[\underline{\theta}^i, \bar{\theta}^i] \subseteq \mathbb{R}$, for $i \in \{M, W\}$. The following observation provides conditions for the existence of equilibria in which each player plays an action with probability 1.

Observation 12. *Under incomplete information, the following statements hold:*
(i) There always exist an equilibrium where Man Yields with probability 1 and Woman Asserts with probability 1, and another equilibrium where it is vice versa.
(ii) There exists an equilibrium where Man and Woman Yield with probability 1*, if and only if* $\underline{\theta}^i \geq 6$*, for* $i \in \{M, W\}$.
(iii) There exists an equilibrium where Man and Woman Assert with probability 1*, if and only if* $\underline{\theta}^i \geq \frac{2}{9}$*, for* $i \in \{M, W\}$.

Note that the thresholds in the observation is the same as those in Observation 11. An implication of this observation is that if all the types are above the complete information threshold, the inefficient equilibrium outcomes (*Y, Y*) and (*A, A*) will be maintained under incomplete information.

Now we study equilibria in which different types play different actions and provide comparative statics results of the acquaintance and stranger societies. In particular, we are interested in the probability of reaching the inefficient outcomes (*Y, Y*) and (*A, A*).

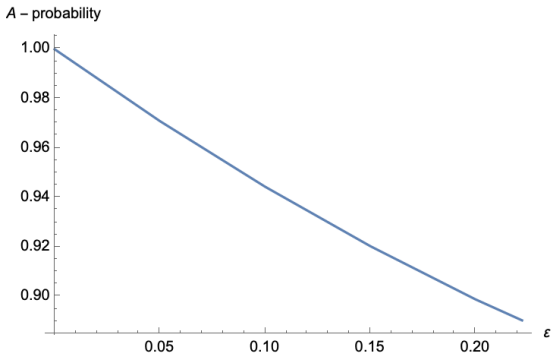


Figure 4: Probability that Each Player Will Play A in Equilibrium with Distribution $Unif[\frac{2}{9}-\varepsilon, 2]$

The previous observation shows that under incomplete information if the players are sure of the opponent's type being greater than 6, (*Y, Y*) can be supported with probability 1. Now, suppose that we assume the type distribution to be $Unif[6-\varepsilon, 50]$ so that while each player is almost certainly of a type above 6, there is a small chance of being a type below 6. One may wonder how the equilibrium would change due to this small change. It turns out that for any $\varepsilon \in (0,5]$, there is no equilibrium in which (*Y, Y*) occurs with any positive probability.

On the other hand, the equilibrium outcome (*A, A*) is robust to a small perturbation. The reason is that, knowing that some types of the opponent might play Y, A becomes materially more attractive for the types playing A that they would not have incentives to deviate to Y. To illustrate, consider type distribution $Unif[\frac{2}{9}-\varepsilon, 2]$ for both players. Figure 4 depicts the probability of each player playing A in equilibrium for different values of ε. Note that the probability of each player playing A in equilibrium decreases continuously as ε increases.

Proposition 4 compares the maximum probability of reaching (*Y, Y*) in the two societies.

Proposition 4. *Suppose that in the stranger society there exists an equilibrium with (Y, Y) occurring with a probability higher than .625. Then, there exists an equilibrium in the acquaintance society with a weakly higher (Y, Y)-probability. The result is strict if* $\underline{\theta}_i < 6$.

In other words, if a society has achieved a high rate of (*Y, Y*) in equilibrium under incomplete information,

it may be *worse* to provide information about each other.[11]

The rough intuition for this result is as follows. Let p_i be the probability of i's choosing Y. First of all, note that the types of Man who are willing to play Y, given that the opponent plays Y (i.e. $p_W = 1$), must be strongly reciprocal (6 or higher in this case). Now, suppose that Woman plays Y with p_W slightly lower than 1. This uncertainty has two effects on Man's utility. First, *Y becomes slightly more profitable* than when $p_W = 1$. Second, since Woman is less unkind, Man's incentive for punishment weakens. As a result *A becomes psychologically more attractive*. For sufficiently high types, the second psychological effect dominates the first material effect. The types who play Y in the acquaintance society ($\theta_i \geq 6$) are reciprocal enough to care about the psychological effect more than the material effect. So, the incentive for punishment in the acquaintance society is higher than that in the stranger society, leading to a higher probability of reaching (*Y, Y*).

4. Revisiting the Prisoners' Dilemma with Reciprocal Motivations

4.1. Equilibrium Characterization

	C	D
C	c,c	$0,x$
D	$x,0$	d,d

Table 1: Prisoners' Dilemma

As has been shown in Rabin (1993), reciprocity concerns can give rise to mutual cooperation in the prisoners' dilemma. When a player cares about not only the material payoff, but also the intention of others, she would think that the co-player is kind by cooperating and unkind by defecting. It is possible that the reciprocal motivation of each player to reward a kind person is so strong that it outweighs the payoff gain from defecting, and then they together can manage to achieve the socially optimal goal through cooperation. The reciprocal equilibrium is characterized by a threshold of the sensitivity parameter. Only when the sensitivity parameters of both players are above this threshold should cooperation happen. Similarly, in the presence of uncertainty, an equilibrium strategy takes the form of a threshold strategy. Nevertheless,

[11] An analogous result does not hold for (*A, A*). In fact, there are many cases in which the maximum probability of (*A, A*) is higher in the stranger society than in the acquaintance society. One example is when the type distributions are $Unif[0,30]$.

the threshold in the incomplete information model is generically different from that in the environment of complete information.

We assume the type spaces to be $\Theta_i = [\underline{\theta}_i, \bar{\theta}_i] \subset \mathbb{R}_+$.[12] Denote by F_i the cumulative distribution function of θ_i which has full support on Θ_i. Furthermore, in this section, we will focus on threshold strategies in the equilibrium characterization. Formally, by a threshold strategy we mean a strategy s_i that has the property that there exists $\theta_i^* \in \mathbb{R}$, such that for $\theta_i \geq \theta_i^*$, $s_i(\theta_i) = C$, and for $\theta_i < \theta_i^*, s_i(\theta_i) = D$.[13] In Proposition 5, we will prove that any equilibrium strategy is a threshold strategy. The reason is that cooperation is strictly dominated by defection with respect to the material payoff, and hence, in order for a player to favor cooperation, the reciprocity payoff from cooperation must be relatively higher than that from defection. Additionally, the payoff is linear in the sensitive parameter, so there will be a lower bound for the types that prefer cooperation.

In the rest of the paper, we simplify the multivariate functions $\kappa_{ij}^{\theta_i}$, $\lambda_{iji}^{\theta_i}$, and $U_i^{\theta_i}$, whose arguments include actions, strategies, and beliefs, into functions that only depend on actions and the *associated* probabilities of strategies. For instance, if under a strategy s_i, player i plays action C with probability 40%, then we say the associated probability of C under s_i is 40%. Specifically, we denote by p_i $(i = 1, 2)$ the probability that player i plays C. In equilibrium, the expression with respect to probabilities is equivalent to the original definition based on two reasons. First, no matter for material payoffs or psychological payoffs, the associated probabilities suffice to pin down players' (expected) payoffs. Second, actions and beliefs coincide at all levels of the belief hierarchy in equilibrium, which allows us to use a single probability to represent both the associated probability of a strategy and the higher-order beliefs about it.

Proposition 5. *Suppose a pair of strategies (s_1, s_2) is an SRE. Then for $i = 1, 2$, s_i takes the form of a threshold strategy.*

Under uncertainty, each player cannot ensure exactly what type the other is assigned and which action the other is taking. Veiled information gives rise to the possibility that one can be kind to an unkind person

[12] In the general model, we assume finite type spaces to facilitate the existence proof of an equilibrium. In the prisoners' dilemma, however, an equilibrium always exists, so we can apply our theory to this case with continuous type space.

[13] For the ease of exposition, we shall assume that any type indifferent between C and D will break the tie in favor of C. This does not affect the equilibrium results in any way, since such a cut-off type only has measure zero.

or be unkind to a kind person. To cooperate brings about a risk of getting betrayed in addition to the material loss, whereas to defect could possibly fail a kind opponent and make the player himself feel bad. Therefore, this paper differs from previous literature by introducing the strategic concern about the innate risks of reciprocating in wrong ways.

Table 1 is a parametric game form of the prisoners' dilemma ($x > c > d > 0$, $2c > x$). We call the row player P1 and the column player P2. In this game, it is a trivial SRE that both players take D at any type, maximizing own material payoffs and reacting to unkindness of each other. To look at a more interesting case, we focus on the "cooperative" SRE that include cooperation with positive probability.

Suppose the players use threshold strategies s_1 and s_2 and p_i $(i = 1,2)$ is the probability that player i takes C. P2's expected payoff ranges from $d(1-p_2)$ to $cp_2 + x(1-p_2)$ depending on the strategy of P1. As the average of the two extremes, the equitable payoff to P2 is $\pi_2^e(p_2) = \frac{1}{2}[cp_2 + d(1-p_2) + x(1-p_2)]$. According to Definition 3, the kindness of P1 to P2 by taking C and D is $\kappa_{12}^{\theta_1}(C,p_2) = \frac{1}{2}[(x-d) + (c+d-x)p_2]$ and $\kappa_{12}^{\theta_1}(D,p_2) = -\kappa_{12}^{\theta_1}(C,p_2)$, respectively. On the other hand, P1 thinks the equitable payoff to herself, symmetric to her opponent, should be $\tilde{\pi}_1^e(p_1) = \frac{1}{2}[cp_1 + d(1-p_1) + x(1-p_1)]$. Hence according to Definition 5, in P1's point of view the kindness of P2 is equal to P1's expected material payoff under strategies s_1 and s_2 minus the equitable payoff $\tilde{\pi}_1^e(p_1)$. It is easy to check that P1 perceives P2's kindness as $\lambda_{121}^{\theta_1}(p_1,p_2) = (p_2 - \frac{1}{2})[(x-d) + (c+d-x)p_1]$.

With these components at hand, based on Eq. (2.3) the utilities of type θ_i from taking C and D can be written as:

$$\begin{aligned} U_i^{\theta_i}(C,p) &= cp_j + \frac{1}{2}\theta_i(p_j - \frac{1}{2})[(x-d) + (c+d-x)p_i] \cdot [(x-d) + (c+d-x)p_j] \\ U_i^{\theta_i}(D,p) &= xp_j + d(1-p_j) - \frac{1}{2}\theta_i(p_j - \frac{1}{2})[(x-d) + (c+d-x)p_i] \cdot [(x-d) + (c+d-x)p_j] \end{aligned} \tag{6}$$

As has been argued before, the equilibrium strategy for player i is featured by a threshold θ_i^*. At θ_i^*, player i must be indifferent between C and D, so that $U_i^{\theta_i^*}(C,p) = U_i^{\theta_i^*}(D,p)$. Solving this equation we have the expression of the threshold $\theta_i^*(p)$ in equilibrium.

$$\theta_i^*(p) = \frac{d - (c+d-x)p_j}{(p_j - \frac{1}{2})[(x-d) + (c+d-x)p_i] \cdot [(x-d) + (c+d-x)p_j]}, \tag{7}$$

where $p_i \neq \frac{1}{2}$ and $(x-d)+(c+d-x)p_i \neq 0$ for $i = 1,2$.[14] Now that Eq. (7) characterizes the threshold strategy for player i, the remaining condition for an equilibrium is that beliefs and strategies should be consistent. That is, for each player i, the proportion of the types above the threshold $\theta_i^*(p)$ according to the original distribution $F_i(\cdot)$ should coincide with his actual cooperation rate p_i under the strategy s_i. Based on Definition 2, the characterization of the SRE in the prisoners' dilemma is as follows.

Proposition 6. *(s_1, s_2) is a cooperative SRE if and only if there is an ordered pair $p = (p_1, p_2) \in (0,1]^2$ such that for each p_i,*

1.
$$s_i(\theta_i) = \begin{cases} C & \text{if } \theta_i \geq \theta_i^*(p), \\ D & \text{if } \theta_i < \theta_i^*(p). \end{cases}$$

2.
$$1 - F_i(\theta_i^*(p)) = p_i. \tag{8}$$

It is worth noting that in any equilibrium that involves cooperation to some extent the associated probability p_j must be strictly higher than one half. Otherwise, player i would view that j cooperates at such a low level that j must be unkind, i.e., $\lambda_{iji}(p_i, p_j) \leq 0$. In this case, player i would have no incentive to cooperate. Thus, player j would not like to cooperate either.

4.2. Stranger vs. Acquaintance Societies

4.2.1. The Condition under which Information Reinforces Cooperation

This section examines the effect of information on cooperation in the prisoners' dilemma by comparing the mutual cooperation rates, the probabilities of achieving the socially optimal outcome (C,C), in the stranger and acquaintance societies. For the purpose of comparisons, We focus on equilibria with the maximal mutual cooperation rate.

[14] When either $p_i = \frac{1}{2}$ or $(x-d)+(c+d-x)p_i = 0$, the psychological term vanishes in the utility function, so that $U_j^{\theta_j}(C,p) = cp_i < xp_i + d(1-p_i) = U_j^{\theta_j}(D,p)$. Player j should play D with probability 1. In turn, player i should play D with probability 1, as well. It contradicts with $p_i = \frac{1}{2}$ or $(x-d)+(c+d-x)p_i = 0$. That means in equilibrium, we do not need to consider these two cases.

In our setup, these two societies differ only in the accessibility of information about sensitivity parameters. Aside from that, players have the same payoff structures and population distributions. It is ambiguous how information asymmetry would influence the mutual cooperation rate. When facing a stranger, a person might be more reluctant to take C considering that her opponent could possibly be mean; but she could also be more willing to take C with the concern that otherwise she might let a kind person down. As we will show below, there is no general answer for this question, but under a certain condition, knowing each other is always conducive to cooperation among players, regardless of their type distributions.

Intuitively, player i's willingness to cooperate should positively correlate with her reciprocity motivation and negatively correlate with the attractiveness of defection. We exhibit this relationship by using the threshold as an indicator; the lower the threshold, the stronger the willingness to cooperate. Then we rewrite Eq. (7) to disentangle the material and reciprocal effects. As in Eq. (9), given that $\theta_i^*(p)$ is positive, all three terms in the fraction are also positive. The numerator is the material gain from defection, which is apparently negatively related to the willingness to cooperate. Meanwhile, the denominator represents the reciprocal payoff to player i when she takes C and it is positively related to i's willingness to cooperate.

$$\theta_i^*(p) = \frac{1}{2} \cdot \frac{\pi_i(D, p_j) - \pi_i(C, p_j)}{\kappa_{ij}^{\theta_i}(C, p_j) \cdot \lambda_{iji}^{\theta_i}(p_i, p_j)} \tag{9}$$

From Eq. (9), the willingness of each player to cooperate depends on her belief about how often the opponent cooperates. In any equilibrium of the acquaintance society, at the moment when they make decisions, they know each other's types and actions. So they will coordinate if both types reach the cutoff level $\theta_i^*(1,1)$ (by symmetry, $\theta_1^*(1,1) = \theta_2^*(1,1)$). Otherwise, both of them will defect. Then, the mutual cooperation rate in the acquaintance society is the probability that both types are above $\theta_i^*(1,1)$, i.e., $[1 - F_1(\theta_1^*(1,1))] \cdot [1 - F_2(\theta_2^*(1,1))]$.

In the stranger society, however, players' beliefs could be anywhere between 0 and 1 and the cutoff level for cooperation is generally different from that in the parallel acquaintance society. To solve for the equilibrium, we make an observation that when $(c+d-x) \geq 0$, θ_i^* is decreasing in p. The number $(c+d-x)$ also equals $c-(x-d)$, which captures the difference in the benefits player i gives j by taking C, conditional on j's choice. When $(c+d-x) \geq 0$, i's taking C is relatively more kind to j when j is taking C. In this

case, if there is a portion of types of j which certainly defect, p_j will be less than one. Not only does cooperation become riskier a choice for i, but from i's point of view j is also less kind to her, which tempers i's enthusiasm to cooperate. Then j will anticipate i's reaction and reduce his cooperation accordingly, which triggers a downward spiral that reaches a lower probability of (C,C). This argument suggests that when c is not too high, in the stranger society the players always achieve lower probability of cooperation.

Proposition 7. *If $c+d-x \geq 0$, the mutual cooperation rate in the acquaintance society is no smaller than that in the stranger society.*

4.2.2. Doubt and Cooperation Breakdown

In the stranger society, lack of information causes doubt among the two players, which could accumulate through iterative deduction and finally reduce or completely break down cooperation. The specific outcome depends on payoff structure and type distributions. To illustrate an extreme case of cooperation breakdown, we propose an example where two persons could very likely cooperate with each other in an acquaintance society, but with no chance in a stranger society.

In this example, the parameters take on values as $c = 2$, $d = 1$, $x = 3$ and it is a special case where $c+d-x = 0$. Then the kindness of each player i from taking C is fixed as 1, while her perceived kindness of j is solely determined by the strategy of j. The threshold becomes a single-variate function, $\theta_i^*(p_j) = 1/2(2p_j - 1)$. Suppose the type of each player is uniformly distributed over $[.45,.95]$. In the acquaintance society, both players could form a cooperative equilibrium (C,C) if their types are above the threshold $\theta^*(1) = \frac{1}{2}$, which accounts for 81% of the time according to the distributions. Strikingly, in the stranger society, cooperation cannot happen at any level. Below we will explain the reason for this sharp contrast.

Initially, player i knows that 10% of the time j will be assigned a type below .5 and will definitely defect. From i's point of view, she is facing this risk for taking C. So not only those types of i below the threshold .5, but also those marginally higher than the cutoff level would like to defect. Specifically, the threshold for i increases to .625 and now with probability 35% she would defect. Taking into account i's thought, j knows he is facing an even bigger risk — he might be failed 35% of the time. Now no type of j from the random draw would like to cooperate, and the same for player i. This process indicates that the suspicion among the two players could loom large until all types retreat from cooperation. The solid curve in Figure 5 illustrates

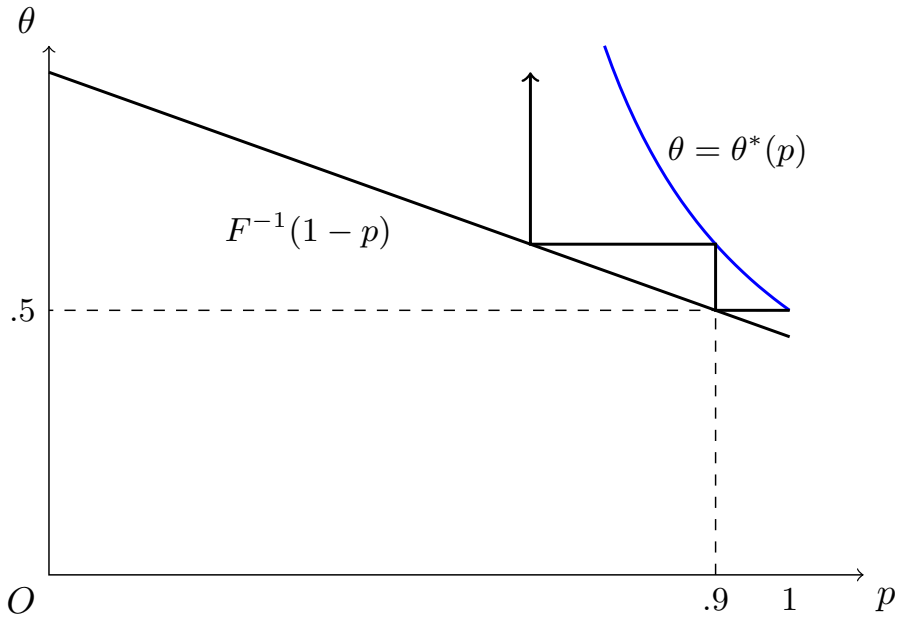


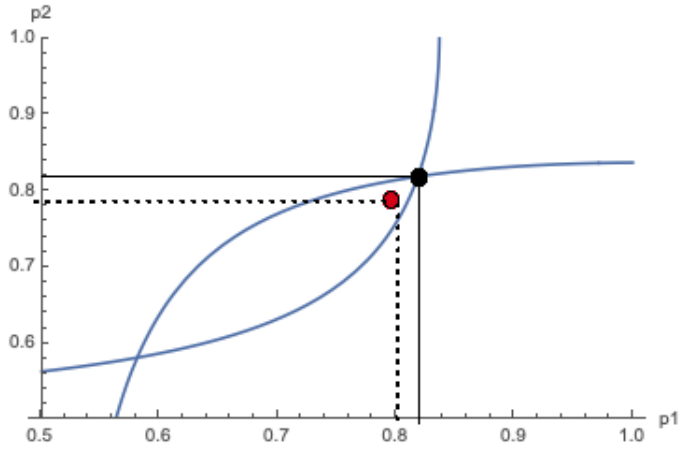
Figure 5: The Collapse of Cooperation

this iterated elimination of cooperative types.

The iterative process is analogous to those in the lemon market (Akerlof, 1970) and in the global game (Carlsson and Van Damme, 1993). To some extent, the sensitive types are prevented from cooperating by growing suspicion in the similar way that quality cars are driven out by lemons and that *risk dominance* is established as the criterion under the unobserved payoff structure. However, these phenomena are driven by different forces. In our model, higher order beliefs come into play through reciprocity payoffs that relate to players' intention; while in the lemon market, the explanation of market breakdown is that the buyer cannot distinguish good cars from bad cars; and in the global game, the beliefs of two players are correlated because of noisy observations of a perturbed game.

4.2.3. When Information Can Hinder Cooperation

In Section 4.2.1, we conclude that, when $c+d-x \geq 0$, information encourages cooperation. If $c+d-x<0$, not having information can be better. We can verify that if $c+d-x<0$, $\theta_i^*(p_i,p_j)$ is increasing in p_i. This makes it possible that for some p_i and p_j, $\theta_i^*(p_i,p_j) < \theta_i^*(1,1)$. So, it is possible that a stranger



Note: The two blue curves refer to Eq. (8) and their intersection indicates the solutions for the equilibrium under incomplete information.

Figure 6: Equilibrium Outcome p_1 and p_2 in Acquaintance Society (Red) and Stranger Society (Black)

society achieves a higher mutual cooperation rate than the acquaintance counterpart, which will be shown in the following example.

	C	D
C	5, 5	0, 9.9
D	9.9, 0	0.1, 0.1

Table 2: Prisoners' Dilemma When $c+d-x<0$

Suppose that the type distribution is uniform on $[0.1, 1.6]$, and the players play the prisoners' dilemma below. In this example, $\theta_i^*(1,1) = .4$. So, in the acquaintance society, players are willing to cooperate if and only if both of their types are greater than or equal to .4. In this case, the acquaintance society achieves mutual cooperation with probability $(.8)^2 = .64$.

In the stranger society, the equilibrium prediction is that a player is willing to cooperate if and only if her type is greater than or equal to 0.372. We can compute that the stranger society achieves mutual cooperation with probability $(.82)^2 \approx .67 > .64$ (See Figure 6). From this example, we know that when the gain from defection, x, is sufficiently small, information will reinforce cooperation; but when x is large, this may not be the case.

To understand why this happens, we should first know that even without information about types, player i knows that with at most 80% probability player j could choose C. Overall, j is still kind to i and i would like to take an action that favors j. It is true that since there is a fraction of types of j that would deviate, from i's perspective, j's kindness is discounted. This causes i's reciprocal motivation to weaken. However, the kindness of i to j by taking C given that j could possibly take D is higher than before, which means i feels better about himself when he chooses to be a generous and forgiving person. These two competing forces determine how information influences players' reciprocity motivations. When the temptation of deviation is large, the increase in the kindness of each player outweighs the decrease in the perceived kindness of their opponents, which contributes to the increase in cooperation under incomplete information.

5. Conclusion

This paper lies in the realm of psychological games (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009). More specifically, it contributes to the literature of reciprocity games (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010). Its innovation is that we assume incomplete information about players' psychological motivations, and we study how introducing uncertainty affects evaluations of kindness and the equilibrium predictions.

We develop a general theoretical framework that can be used to study reciprocity games in extensive-form games with incomplete information. In our framework, players update their beliefs about other players' types at each stage, which would further change their understandings of others' intentions. Therefore, information plays an important role in determining the level of kindness and the play of the game.

We apply our model to a series of well known games and obtain several insightful results. We also analyze the comparative statics between acquaintance and stranger societies. Take the prisoners' dilemma for example. In standard game theory, it is the unique Nash equilibrium in which both players choose to defect. Then Rabin (1993) shows that mutual cooperation can actually be supported as a reciprocal equilibrium when both players are sufficiently sensitive. However, we show that cooperation may not be robust to a low probability that both players might be relatively insensitive. Furthermore, the cooperation rate is generally higher in the acquaintance society than in the stranger society, when the benefits of defection are modest.

References

Akerlof, G.A., 1970. The market for "lemons": Quality uncertainty and the market mechanism. The quarterly journal of economics , 488–500.

Andreoni, J., 1993. An experimental test of the public-goods crowding-out hypothesis. The American Economic Review , 1317–1327.

Antler, Y., 2015. Two-sided matching with endogenous preferences. American Economic Journal: Microeconomics 7, 241–258.

Attanasi, G., Battigalli, P., Manzoni, E., 2016. Incomplete-information models of guilt aversion in the trust game. Management Science 62, 648–667.

Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. Journal of Economic Theory 144, 1–35.

Bellemare, C., Sebald, A., Suetens, S., 2018. Heterogeneous guilt sensitivities and incentive effects. Experimental Economics 21, 316–336.

Bierbrauer, F., Netzer, N., 2016. Mechanism design and intentions. Journal of Economic Theory 163, 557–603.

Carlsson, H., Van Damme, E., 1993. Global games and equilibrium selection. Econometrica: Journal of the Econometric Society , 989–1018.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., 2008. Representative trust and reciprocity: Prevalence and determinants. Economic Inquiry 46, 84–90.

Dufwenberg, M., Gachter, S., Hennig-Schmidt, H., 2011. The framing of games and the psychology of play. Games and Economic Behavior 73, 459–478.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and economic behavior 47, 268–298.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and economic behavior 54, 293–315.

Fudenberg, D., Tirole, J., 1991. Perfect bayesian equilibrium and sequential equilibrium. Journal of Economic Theory 53, 236–260.

Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games and economic Behavior 1, 60–79.

Gronberg, T.J., Luccasen III, R.A., Turocy, T.L., Van Huyck, J.B., 2012. Are tax-financed contributions to a public good completely crowded-out? experimental evidence. Journal of Public Economics 96, 596–603.

Hennig-Schmidt, H., Sadrieh, A., Rockenbach, B., 2010. In search of workers' real effort reciprocity—a field and a laboratory experiment. Journal of the European Economic Association 8, 817–837.

Kocher, M.G., Martinsson, P., Persson, E., Wang, X., 2016. Is there a hidden cost of imposing a minimum contribution level for public good contributions? Journal of Economic Psychology 56, 74–84.

Kozlovskaya, M., Nicolo, A., 2019. Public good provision mechanisms and reciprocity. Journal of Economic Behavior and Organization 167, 235–244.

Kreps, D.M., Wilson, R., 1982. Sequential equilibria. Econometrica: Journal of the Econometric Society , 863–894.

Martinsson, P., Medhin, H., Persson, E., 2019. Minimum levels and framing in public good provision. Economic Inquiry 57, 1568–1581.

Rabin, M., 1993. Incorporating fairness into game theory and economics. The American economic review , 1281–1302.

Sebald, A., 2010. Attribution and reciprocity. Games and Economic Behavior 68, 339–352.

Van Damme, E., Binmore, K.G., Roth, A.E., Samuelson, L., Winter, E., Bolton, G.E., Ockenfels, A., Dufwenberg, M., Kirchsteiger, G., Gneezy, U., et al., 2014. How werner güth's ultimatum game shaped our understanding of social behavior. Journal of economic behavior & organization 108, 292–318.

		Bob	
		C	D
Ann	C	$(3,3)$	$(0,5)$
	D	$(5,0)$	$(1,1)$

Figure 7: Prisoner's Dilemma

Appendix A

A.1 Alternative Definitions of Kindness and Perceived Kindness

Kindness could have been defined differently from the definition we proposed. Here we study one alternative definition of kindness and one alternative definition of perceived kindness, and discuss why the alternative definitions are not suitable. Instead of defining kindness in terms of a specific action taken by the realized type, we could have defined kindness in terms of the strategy. Under this alternative definition, kindness of a player is unaffected by the realized type. Formally, the following would be the alternative formulation of kindness of Player i with type θ_i to player j.

$$\tilde{\kappa}_{ij}(a_{i,h},(b_{ij,h})_{j\neq i},\mu(h)) = \pi_j(a_{i,h},(b_{ij,h})_{j\neq i},\mu(h)) - \tilde{\pi}^e_{ij}((b_{ij,h})_{j\neq i},\mu(h)),$$

where the equitable payoff $\tilde{\pi}^e_{ij}$ is defined as follows:

$$\tilde{\pi}^e_{ij}((b_{ij,h})_{j\neq i},\gamma) = \frac{1}{2}[\max\{\pi_j(a_i,(b_{ij,h})_{j\neq i}),\gamma)|a_i\in A_i\} + \min\{\pi_j(a_i,(b_{ij,h})_{j\neq i}),\gamma)|a_i\in \tilde{E}_i\}]$$

The idea here is that the player's kindness could have been defined as a function of the whole strategy rather than the action he ends up playing. Let us explain how this alternative definition is not suitable with an example. Consider an example of the Prisoners' Dilemma in Figure 7. First of all, under complete information, note that D is an unkind action and C is a kind action regardless of one's belief about the action of the other.

$$a_{Ann}(\theta_{Ann}) = \begin{cases} \text{D} & \text{if } \theta_i = .50 \\ \text{C} & \text{Otherwise} \end{cases} \tag{10}$$

Suppose each player's type is drawn from the set $\{0.01,0.02,\cdots,0.99,1\}$ with equal probabilities. Ann's

		Bob	
		F	B
Ann	F	$(1,3)$	$(0,0)$
	B	$(0,0)$	$(3,1)$

Figure 8: battle of the sexes (BoS)

strategy is as in Eq. (**??**). Say the realized type is $\theta'_{Ann} = 0.50$. Following the strategy, Ann plays D. Is Ann kind? We may think that θ_{Ann} is Ann's exogenously given personal trait and is determined before the game which is unknown to others. It seems natural that Ann should consider herself unkind since Ann is playing D, given the type. If Ann considered what she could have done (C with probability .99) instead of what she happens to do (D), then she could consider herself as being kind. But, it seems implausible that people perceive themselves as kind due to an action that they did not do, but *hypothetically could have done*. Another reason is a mathematical one. Suppose that Ann evaluates her kindness over all her possible types, according to the alternative definition. Then, Ann's action given the type does not affect her kindness much, because her realized type has a very small weight on 0.50. This implies that the reciprocity payoff term becomes inconsequential for most types.

Type θ_i of player i's perception about j's kindness could also be defined in a different way. Recall that type θ_i's perception about j's kindness was about how kind j is to i's *ex ante* payoff. Alternatively, θ_i's perception about j's kindness could have been about j's kindness to the realized type of i, θ_i. Now, we give an example to illustrate why we should favor our definition, by arguing against the alternative definition. Let $c^{\theta_i}_{iji,h}$ refer to i's second order belief about his own action for type θ_i. Consider the following alternative definition, which is a formalization of the alternative idea:

$$\tilde{\lambda}^{\theta_i}_{iji}(b_{ij,h},(c_{ijk,h})_{k\neq j},\mu_{-i}(h)) = \pi_i(c^{\theta_i}_{iji,h},b_{ij,h},(c_{ijk,h})_{k\neq j,i},\mu_{-i}(h)) - \tilde{\pi}^{e}_{iji}((c_{ijk,h})_{k\neq j},\mu_{-i}(h))$$

where the equitable payoff is defined as follows:

$$\tilde{\pi}^{e}_{iji}((c_{ijk,h})_{k\neq j},\mu_{-i}(h)) = \frac{1}{2}[\max\{\pi_i(c^{\theta_i}_{iji,h},a_j,(c_{ijk,h})_{k\neq j,i},\mu_{-i}(h))|a_j\in A_j\} + \min\{\pi_i(c^{\theta_i}_{iji,h},a_j,(c_{ijk,h})_{k\neq j,i},\mu_{-i}(h))|a_j\in \tilde{E}_j\}].$$

Now, consider the battle of the sexes (BoS) in Figure 8 between Ann and Bob. Their types have the same

probability distributions as the previous example. Ann and Bob's strategies are specified as below:

$$a_{Ann}(\theta_{Ann}) = \begin{cases} \text{F} & \text{if } \theta_i = 0.50 \\ \text{B} & \text{Otherwise} \end{cases}$$

$$a_{Bob}(\theta_{Bob}) = \text{B for all } \theta_{Bob}$$

Suppose the types of the players happen to be $\theta_{Ann} = \theta_{Bob} = 0.50$. Should Ann think of Bob as unkind? Ann is playing F in this example, and Bob is playing B with probability 1. Using our alternative formulation, it may seem that Bob is indeed unkind to Ann, as Ann's type is 0.50. However, this implies that Ann holds Bob accountable for what Bob does not know. Bob clearly does not know Ann's type as it is private information, and the strategy dictates that Ann is playing B with probability .99. Thus, the best action Bob could have chosen in order to help Ann was indeed B, without knowing the type of Ann. If Ann could think this through, Ann should not hold Bob accountable, since Bob has done the best action he could, given Bob's information. The alternative definition leads to psychologically incoherent implications.

A.2 Proof of Theorem 1

Proof. Let us first define a perturbed game $\Gamma(\varepsilon)$, where players are restricted to play completely mixed strategies. That is, for any θ_i, h and $d \in D_i(h)$, player i's strategy satisfies that $a_i(\theta_i, h)(d) \geq \varepsilon$, and ε is sufficiently small so that $|D_i(h)| \cdot \varepsilon \leq 1$.

We examine a sequence of perturbed games $\Gamma(\varepsilon_n)$ with $\varepsilon_n \to 0$. In each perturbed game $\Gamma(\varepsilon_n)$, the belief system is uniquely determined by a strategy profile a according to Bayes rule, written as μ_a.

First we prove that in a perturbed game, an equilibrium assessment exists. The proof resembles that of existence of reciprocity equilibrium in complete information games in DK (2004). Since in a perturbed game the belief system is a continuous function of the strategy profile, it suffices to find a fixed point of the self-mapping on A. The existence of a fixed point is achieved by the standard applications of the Maximum Theorem and Kakutani Theorem. Note that the conditions for these theorems are satisfied. First, $U_i^{\theta_i}$ is continuous in the behavior strategy, and the first-, and second-order beliefs, so the Maximum Theorem applies. Second, $U_i^{\theta_i}$ is linear in s_i, thus the best-response correspondence is convex-valued, and Kakutani

theorem applies.

Next, because A is compact, we can select a subsequence of $\{a^n\}$ such that a^n is an SRE in $\Gamma(\varepsilon_n)$, and $\mu^n = \mu_{a^n} \longrightarrow \mu^*$ and $a^n \longrightarrow a^*$.

At last, we show that (a^*, μ^*) is an equilibrium assessment of the original game. By definition, (a^*, μ^*) is consistent as the limiting point of a sequence of assessments under completely mixed strategies. What remains to be shown is that each agent (i,h) maximizes his utility. Suppose not, then there is a type who can deviate at history h from $a^*_{i,h}(\theta_i)$ to $s_i \in S_i(\theta_i, h, a^*)$ and

$$U_i^{\theta_i}(s_i, (b^*_{ij,h}, (c^*_{ijk,h})_{k\neq j})_{j\neq i}, \mu^*(h)) > U_i^{\theta_i}(a^*_{i,h}(\theta_i), (b^*_{ij,h}, (c^*_{ijk,h})_{k\neq j})_{j\neq i}, \mu^*(h)) \tag{11}$$

where $b^*_{ij,h} = a^*_{j,h}$ and $c^*_{ijk,h} = b^*_{jk,h}$, for all i, j, k.

Then we construct a sequence of strategies $s_i^n \in S_i$ that satisfy three conditions: (1) s_i^n is a completely mixed strategy in $\Gamma(\varepsilon_n)$; (2) $s_i^n = a^n_{i,h'}(\theta_i)$ for all history h' except h; (3) s_i^n converges to s_i. Because $U_i^{\theta_i}$ is continuous, when n is large enough, $U_i^{\theta_i}(s_i^n, (b^n_{ij,h}, (c^n_{ijk,h})_{k\neq j})_{j\neq i}, \mu^n(h))$ approximates $U_i^{\theta_i}(s_i, (b^*_{ij,h}, (c^*_{ijk,h})_{k\neq j})_{j\neq i}, \mu^*(h))$ and $U_i^{\theta_i}(a^n_{i,h}(\theta_i), (b^n_{ij,h}, (c^n_{ijk,h})_{k\neq j})_{j\neq i}, \mu^n(h))$ approximates $U_i^{\theta_i}(a^*_{i,h}(\theta_i), (b^*_{ij,h}, (c^*_{ijk,h})_{k\neq j})_{j\neq i}, \mu^*(h))$. Because of Ineq. (11), there is n such that

$$U_i^{\theta_i}(s_i^n, (b^n_{ij,h}, (c^n_{ijk,h})_{k\neq j})_{j\neq i}, \mu^n(h)) > U_i^{\theta_i}(a^n_{i,h}(\theta_i), (b^n_{ij,h}, (c^n_{ijk,h})_{k\neq j})_{j\neq i}, \mu^n(h))$$

where $b^n_{ij,h} = a^n_{j,h}$ and $c^n_{ijk,h} = b^n_{jk,h}$, for all i, j, k. That means that in the perturbed game $\Gamma(\varepsilon_n)$ after history h, the type θ_i of player i can profitably deviate to s_i^n, which contradicts that (a^n, μ^n) is an equilibrium assessment in $\Gamma(\varepsilon_n)$. □

Appendix B

B.1 Investment Game with Punishment

Proof of Observation 1. When $p \geq \frac{1}{3p}$, Ineq. (1) also holds. That means that after reaching (*Invest*, *Grab*), the reciprocal type of P1 would prefer *Punish*, which can deter P2 from choosing *Grab*. Given that P2 would like to *Share*, the reciprocal type of P1 would take *Invest* according to Ineq. (2). □

B.2 Ultimatum Game

Proof of Proposition 1. There are a proposer (P) and a responder (R). I will use male noun (he) for P and female noun (she) for R. P offers a split of a unit pie into $(1-x,x)$, $x \geq 0$, and R decides to *accept* or *reject*. If R accepts the offer, she will receive a payment x, and P will receive $1-x$. Otherwise, they get zero payoffs.

P's pure strategy is a number $x \in [0,1]$. R's pure strategy is a mapping $s : [0,1] \to \{A,R\}$. Suppose that R plays a threshold strategy with a cutoff level $\bar{s}$. Under this strategy, she will accept if $x \geq \bar{s}$ and vice versa. For simplicity and without loss of the insight, we focus on the case where P is purely selfish and R accepts the offer on the equilibrium path. Given that there is a continuum of equilibria, we only study P's favorite equilibrium and make comparative statics analysis.[15]

When P gives R an offer $x = 1$, R will unarguably accept it due to the maximal material payoff and kindness of P. That means if P offers $x = 0$, he must be mean to R and R has every reason to reject. With this in mind, we know that the equitable payoff to R is $\frac{1}{2}$.

If P gives an offer x under the expectation that R will accept it, then P's kindness to R is $\kappa_P = x - \frac{1}{2}$. In face of an offer $x > 0$,[16] the only efficient strategy for R is *Accept*, therefore, the equitable payoff to P is $1-x$ and the kindness of R by accepting the offer is 0, i.e., $\kappa_R = 0$.

From above, R's utility from accepting the offer is x and that from rejecting the offer is $-\frac{\theta_R}{2}(1-x)(2x-1)$. When $x \in [0, \frac{1}{2}]$, the former is increasing and the latter is decreasing in x; when $x \geq \frac{1}{2}$, the former is positive while the latter is negative. These two facts justify that an equilibrium strategy takes a form of threshold strategy. At the cutoff level $x = \bar{s}$, R must be indifferent. After transformation, $\bar{s}$ satisfies

$$2\bar{s}^2 - (3 + \frac{2}{\theta_R})\bar{s} + 1 = 0 \tag{12}$$

One of the two solutions to this equation is out of the range $[0,1]$, so we only consider the other reasonable

[15] The reason that there could be multiple cutoff levels for equilibrium lies in that the kindness of P depends on his expectation of R's response. If he gives a positive offer which he expects R to accept, it would be nice of him; but for the same offer, if he expects it to be rejected, then this move should be regarded as unkind. Therefore, it is possible that a positive offer is chosen by P and accepted by R in one equilibrium but rejected in another equilibrium.

[16] In concern with equilibrium outcomes, we only need to study the case of $x > 0$. An offer $x = 0$ can never be accepted in equilibrium, because in terms of material payoff R is indifferent, but in terms of psychological payoff she wants to punish P's unkindness.

solution.

$$\bar{s}(\theta_R) = \frac{1}{4}\left[(3+\frac{2}{\theta_R}) - \sqrt{(3+\frac{2}{\theta_R})^2 - 8}\right] \tag{13}$$

The first-order derivative is

$$\frac{\partial \bar{s}}{\partial \theta_R} = \frac{1}{4}\left(3+\frac{2}{\theta_R}\right)' - \frac{1}{4}\frac{\left(3+\frac{2}{\theta_R}\right)\left(3+\frac{2}{\theta_R}\right)'}{\sqrt{\left(3+\frac{2}{\theta_R}\right)^2 - 8}}$$

Because $\left(3+\frac{2}{\theta_R}\right)'$ is negative and $\left(3+\frac{2}{\theta_R}\right) > \sqrt{\left(3+\frac{2}{\theta_R}\right)^2 - 8}$, so $\partial\bar{s}/\partial\theta_R > 0$ for all $\theta_R \in (0,+\infty)$. Furthermore, when $\theta_R \to +\infty$, $\bar{s} \to \frac{1}{2}$; when $\theta_R \to 0$, $\bar{s} \to 0$. Therefore, the solution functions have nice monotonicity and asymptotic properties. □

B.3 Monopoly Pricing

Proof of Observation 2. Consider any equilibrium where C chooses the reservation price, r, which means that for any price $p > r$, C is happy to refuse to buy from the monopolist, and for any price $p \leq r$, C is happy to buy.

Then, we can derive the following conditions:

$$p \leq r \Rightarrow (v-p) + \theta(p-c)((v-p) - \frac{1}{2})) \geq 0 \tag{14}$$

$$p > r \Rightarrow (v-p) + \theta(p-c)(\eta(v-p) - \frac{1}{2})) < 0. \tag{15}$$

Therefore, any equilibrium reservation price, r, must satisfy the inequalities above simultaneously. Let $x = v-p$. $1-x = p-c$. Then, Eq. (14) can be written as $x + \theta(1-x)(x-\frac{1}{2}) \geq 0$, and $x \in [0,1]$.

By solving this inequality, the solution set is calculated to be $[\frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4}, \frac{(3+2/\theta)+\sqrt{(3+2/\theta)^2-8}}{4}]$. But, it can be easily checked that while $x \leq 1$, $\frac{(3+2/\theta)+\sqrt{(3+2/\theta)^2-8}}{4} > \frac{3+\sqrt{(3)^2-8}}{4} = 1$. So, the set of solutions can be written as $x \in [\frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4}, 1]$. (It can also be checked that $\frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4} \in (0,1)$.)

Similarly, Eq. (15) Can be written as $x + \theta(1-x)(\eta x - \frac{1}{2}) < 0$. By solving this inequality we get the set

of solutions as $x < \frac{(2\eta+1+2/\theta)-\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta}$ or
$x > \frac{(2\eta+1+2/\theta)+\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta}$. Now, I show that $\frac{(2\eta+1+2/\theta)+\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta} > 1$, and thus, there is no value of $x \in [0,1]$ that satisfies this.

$\frac{(2\eta+1+2/\theta)+\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta} > \frac{(2\eta+1)+\sqrt{(2\eta+1)^2-8\eta}}{4\eta} = \frac{(2\eta+1)+\sqrt{(2\eta-1)^2}}{4\eta} = 1$. Hence, the solutions of Eq. (15) can be written as $x \in [0, \frac{(2\eta+1+2/\theta)-\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta})$.

Now, we show that $\frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4} < \frac{(2\eta+1+2/\theta)-\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta}$. Suppose it weren't true. If $\frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4} > \frac{(2\eta+1+2/\theta)-\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta}$, then there is a value in between such that neither Eq. (14) nor Eq. (15) are not satisfied. But, by studying Eq. (14) and Eq. (15) it is clear that at least one of the inequalities must be satisfied. So, this only possibility we need to consider is when $\frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4} = \frac{(2\eta+1+2/\theta)-\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta}$. Since Eq. (14) holds with equality at this value, Eq. (15) must hold with inequality. Then, there is a value $\alpha > \frac{(2\eta+1+2/\theta)-\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta}$, such that Eq. (15) holds, which contradicts our statement that $x < \frac{(2\eta+1+2/\theta)-\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta}$ characterizes the set of solutions of Eq. (15). □

Proof of Observation 3. $\frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4} < \frac{1}{2}$. This can be shown by showing that the derivative of $\frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4}$ w.r.t θ_C is positive, and the limit of it as $\theta_C \to \infty$ is $\frac{1}{2}$. So, There always exists an reservation price such that $1-x > 1-\frac{1}{2} \geq \eta$, that can be supported in equilibrium. And since $\eta < 1-x$, M will offer the reservation price at which both types will buy. □

Observation. *When the highest-reservation-price equilibria are compared, if $\eta \leq \frac{1}{2}$, the price offered in the stranger society is identical to the price offered to the reciprocal type in the acquaintance society.*

Proposition 3-(i). Consider the stranger society. We have already shown that if $\eta \leq 1/2$ there is some price $p < v$ offered by M in equilibrium. In fact, we showed this by showing that there is an equilibrium where M offers $\bar{p} = v - \frac{(3+2/\theta)-\sqrt{(3+2/\theta)^2-8}}{4}$, and this is the highest possible reservation price in equilibrium. In this equilibrium, both types buy at the given price.

In the acquaintance society, the reservation price for the reciprocal type must satisfy the following inequalities:

$$p \leq r \Rightarrow (v-p) + \theta(p-c)((v-p) - \frac{1}{2})) \geq 0 \tag{16}$$

$$p > r \Rightarrow (v-p) + \theta(p-c)(-\frac{1}{2})) < 0. \tag{17}$$

The solution set of the inequalities is: $[\bar{p}, \underline{\underline{p}}]$, where $\underline{\underline{p}} = v - \frac{\theta}{2+\theta}$.

Note that the highest reservation price is $\bar{p}$ in both societies. In the acquaintance society, it is the monopolist's best response to offer the price that match the reservation price to reap a positive profit. So, it constitutes an equilibrium.

As a result, when the highest-reservation-price equilibria are compared, if $\eta \leq \frac{1}{2}$, the price offered in the stranger society is identical to the price offered to the reciprocal type in the acquaintance society. □

Observation. *(Observation 5-(ii)) When the lowest-reservation-price equilibria are considered, M's equilibrium price is strictly higher in the stranger society than the price for the reciprocal consumer in the acquaintance society.*

Proposition 3-(ii). In the stranger society, the lowest reservation price is $\underline{p} := v - \frac{(2\eta+1+2/\theta)+\sqrt{(2\eta+1+2/\theta)^2-8\eta}}{4\eta}$, which can be seen from the proof of Proposition 3-(i). Thus, the monopolist either offers price $p = v$ or $p = \underline{p}$. In the acquaintance society, the monopolist matches the reservation price at $\underline{\underline{p}}$. And, as we have shown previously, $\underline{\underline{p}} < \underline{p} < v$. □

B.4 Sequential Prisoners' Dilemma

Suppose player i's sensitivity parameter is a random variable that would take on values either 0 or θ_i (> 0). The prior probability of θ_i is p_i. Then we examine under what conditions should the cooperative equilibrium, where the unselfish types of both players 100% choose C, be supported.

In the second round, conditional on the decision node following D, P2 revises her belief about P1's strategy such that she would treat P1 as if he was playing a strategy that assigns probability 1 to D. Based on this belief, she would view P1 as unkind, and choose d irrespective of her type.

On the contrary, if the decision node following C is reached, P2, when she is unselfish, may want to choose c when she is strongly motivated by reciprocity. If she does choose c, her kindness to P1 is 1, because she can either give P1 2 or 0. If her choice is d, then her kindness becomes -1. Given that we are interested in equilibria where P2 plays strategy (c,d) when she is unselfish and (d,d) when she is selfish, the highest

expected payoff for P2 is obtained when P1 chooses C, which equals $3-p_2$; the lowest is when P1 chooses D, which equals 1. Thus, P2 would perceive P1's kindness of choosing C as $\frac{1}{2}[(3-p_2)-1] = 1-\frac{1}{2}p_2$.

To support the choice of c, it needs to be satisfied that:

$$2+\theta_2(1-\frac{1}{2}p_2) \geq 3+\theta_2(-1)(1-\frac{1}{2}p_2)$$

It can be simplified to the condition $\theta_2 \geq \frac{1}{2-p_2}$. Note that the threshold is lower than that in DK (2004), which means that it is easier for P2 to cooperate. That is because P2 acknowledges the risk on P1 when he chooses C that she might be selfish and would choose d in response. So she is more grateful for P1's generosity when he truly chooses C.

Then, we discuss P1's equilibrium strategy. There are two cases according to whether the selfish type of P1 cooperates. By considering material expected payoffs, it is easy to see that the unselfish type of P1 would choose C if $p_2 \geq \frac{1}{2}$ and D if $p_2 < \frac{1}{2}$. Let's say the first case is *Case* 1 and the second case is *Case* 2.

In both cases, P1's material expected payoffs and kindness are the same. If he takes C, his material expected payoff equals $2p_2$ and kindness $1-\frac{1}{2}p_2$; if he takes D, his material expected payoff equals 1 and kindness $\frac{1}{2}p_2-1$. The only difference is concerned with the perceived kindness. In Case 1, given that P2 plays the equilibrium strategy of interest, P1 would view her kindness as being $2p_2-1$.[17] In Case 2, the perceived kindness equals $2p_1-1$.[18]

With these numbers at hand, we are able to calculate the condition under which the unselfish type of P1 is willing to cooperate.

In Case 1, the condition is

$$2p_2+\theta_1(1-\frac{1}{2}p_2)(2p_2-1) \geq 1+\theta_1(\frac{1}{2}p_2-1)(2p_2-1)$$

[17] Since P1 chooses C regardless of his types, P2's kindness is calculated as if P1 played a pure strategy in a complete information game.

[18] The maximum payoff for P1 would be $2p_1+3(1-p_1)$, and the minimum payoff for P1 would be $(1-p_1)$. The fair payoff is $2-p_1$. The actual payoff is $2p_1+(1-p_1)$. Therefore, the perceived kindness is $2p_1-1$.

Transforming the above inequality, we obtain

$$(2p_2-1)+\theta_1(2-p_2)(2p_2-1)\geq 0$$

Note that in Case 1 we have $p_2 \geq \frac{1}{2}$, therefore the inequality always holds.

In Case 2, the condition is

$$2p_2+\theta_1(1-\frac{1}{2}p_2)(2p_1-1)\geq 1+\theta_1(\frac{1}{2}p_2-1)(2p_1-1)$$

If $p_1 \leq \frac{1}{2}$, the inequality does not hold. If $p_1 > \frac{1}{2}$, it holds when

$$\theta_1 \geq \frac{1-2p_2}{(2p_1-1)(2-p_2)}$$

B.5 Public Goods Game

Proof of Observation 7. W.L.O.G, consider Player 1's incentive. The material payoff functions is

$$\pi_1 = 20-s_1+\frac{1}{2}[s_1+s_2+s_3].$$

The kindness function can be computed as follows:

$$\begin{aligned}\kappa_{12}(s_1,b_{12},b_{13}) = &\quad \pi_2(s_1,b_{12},b_{13})-\tfrac{1}{2}[\max_{s_1'}\{\pi_2(s_1')\}+\min_{s_1'}\{\pi_2(s_1')\}]\\ = &\quad \tfrac{1}{2}(s_1+\mathbb{E}[b_{12}]+\mathbb{E}[b_{13}])-\tfrac{1}{2}[10+\tfrac{x}{2}+\mathbb{E}[b_{12}]+\mathbb{E}[b_{13}]].\\ = &\quad \tfrac{1}{2}[s_1-\tfrac{20+x}{2}].\end{aligned}$$

Analogously, we can show that the perceived kindness term is computed as:

$$\lambda_{1j1} = \frac{1}{2}[\mathbb{E}[b_{1j}]-\frac{20+x}{2}].$$

Therefore, the utility from contributing s_1 is

$$U_1(s_1,b_{12},b_{13}) = 20 - s_1 + \frac{1}{2}(s_1+s_2+s_3) + \frac{\theta_i}{2}[s_1 - \frac{20+x}{2}][\mathbb{E}[b_{12}] + \mathbb{E}[b_{13}] - 20 - x].$$

Note that this expression is linear in s_1. This means that either the player contributes all or 0, indifferent in the level of contribution. $\partial_{s_1} U_1 \geq 0 \Leftrightarrow \theta_1[\mathbb{E}[b_{12}] + \mathbb{E}[b_{13}] - 20 - x] \geq 2$. □

Proof of Observation 8. (i) Suppose that $p \leq 1/2$. This means that at least half of the population is selfish enough to contribute 0. Therefore, $[\mathbb{E}[b_{1j}] \leq 10$. So, $\theta_1[\mathbb{E}[b_{12}] + \mathbb{E}[b_{13}] - 20 - x] \leq -x(\theta_1) < 2$. So, Player 1 will not contribute anything regardless of the type. By symmetry, no player contributes anything.

(ii) Suppose that $\bar{\theta}_i \geq \frac{2}{20(2p-1)-x}$ or $\bar{\theta}_i(20(2p) - 20 - x) \geq 2$. It can be easily check that all the high types are willing to fully contribute when the high types of others also fully contribute.

□

Proof of Observation 9. Since there is a player whose high type is $\bar{\theta}_i \in [\frac{2}{20(2p-1)}, \frac{2}{20(2p-1)-x})$, this player will not contribute anything when there is a tax of x. Accordingly no one would contribute, except the tax. So, the total contribution is $3x$. Now, suppose that there is no tax at all. Then, since all of the high types are above $\frac{2}{20(2p-1)}$, they will contribute fully, when there is no tax. So, the society achieves $60p$ as the expected total contribution. $60p > 3x$, since $20p > x$. □

Proof of Observation 10. Suppose that $p > \frac{1}{2}$ and $\theta_i \geq \frac{2}{20(2p-1)-x})$. Then, we know that in the stranger society, there is an equilibrium where all the high types contribute fully. Hence, the expected contribution is $60p + 3(1-p)x$. In the acquaintance society, the high type of each player contributes only when all other types are of high types. Thus, the expected contribution level is $60p^3 + 3(1-p^3)x$, which can be shown to be less than $60p + 3(1-p)x$. □

B.6 battle of the sexes

While keeping in mind the symmetry of the game, consider Man's problem. Let p^M and p^W be the probabilities of Yield of Man and Woman. Man's utility from each action in equilibrium is computed as follows:

$$U_M^{\theta^M}(Yield, p_M, p_W) == (1-p_W) + \theta^M[\frac{3}{2} - 2p_W][2p_M + 3p_W - 4p_m p_W - \frac{3}{2}]$$

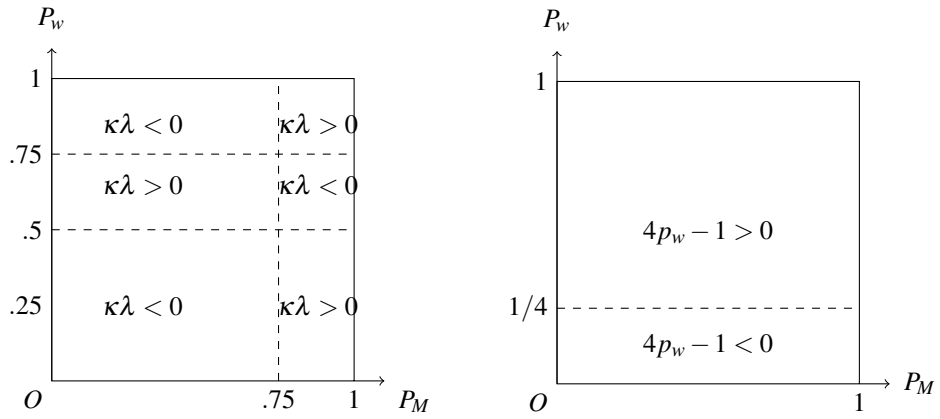


Figure 9: The Sign of $\kappa\lambda$ and $4p_w - 1$

$$U_M^{\theta^M}(Assert, p_M, p_W) = (1 - p_W) + \theta^M[-\frac{3}{2} + 2p_W][2p_M + 3p_W - 4p_m p_W - \frac{3}{2}]$$

By comparing the equations, it can be inferred that θ^M will prefer Yield (Assert) if

$$2\theta^M[\frac{3}{2} - 2p_W][2p_M + 3p_W - 4p_m p_W - \frac{3}{2}] \geq (<) 4p_W - 1. \tag{18}$$

Therefore, any type θ^M that satisfies the inequality strictly must prefer Yield (Assert). Denote by $\kappa := \frac{3}{2} - 2p_W$, and by $\lambda := 2p_M + 3p_w - 4p_M p_W - \frac{3}{2}$, so the inequality can simply be written as $2\theta^M\kappa\lambda \geq 4p_W - 1$. The sign of $\kappa\lambda$ is depicted in Figure 9.

Remark 2. *If the values of* (p_M, p_W) *lie in* $[0, 0.75] \times [0.75, 1]$, $[0.75, 1] \times [0.5, 0.75]$, *or* $[0, 0.75] \times [0.25, 0.5]$, *Man of every type* (> 0) *strictly prefers Assert, and if* p_M *and* p_W *lie in* $[0.75, 1] \times [0, 0.25]$, *Man of every type* (> 0) *prefers Yield, except at* $(p_M, p_Y) = (\frac{3}{4}, \frac{1}{4})$ *where all types are indifferent between the two actions.*

Proof. Suppose that (p_M, p_W) lie in $[0, 0.75] \times [0.75, 1]$ or $[0.75, 1] \times [0.5, 0.75]$. Then, $\kappa\lambda \leq 0$ and $4p_W - 1 > 0$.

So, for all $\forall\theta^M \geq 0$, $2\theta^M\kappa\lambda \leq 0 < 4p_W - 1$. θ^M strictly prefers Assert.

Now, suppose that (p_M, p_W) lie on the northern and eastern boundaries. Then, $\kappa\lambda = 0$ and $4p_W - 1 > 0$.

So, $2\theta^M\kappa\lambda = 0 < 4p_W - 1$. $\Rightarrow \theta^M$ strictly prefers Assert.

Now, suppose that (p_M, p_W) lie on the southern border. Then, $4p_W - 1 = 0$ and $\kappa\lambda < 0$. So, $2\theta^M\kappa\lambda < 0 = 4p_W - 1$. $\Rightarrow \theta^M$ strictly prefers Assert. And if (p_M, p_W) lie in the interior of $[0, 0.75] \times [0.25, 0.5]$, $2\theta^M\kappa\lambda < 0 < 4p_W - 1$.

Similarly, in $[0.75, 1] \times [0, 0.25]$, $2\theta^M\kappa\lambda > 0 \geq 4p_W - 1$ (except at $(\frac{3}{4}, \frac{1}{4})$ where both $\kappa\lambda = 0$ and $4p_W - 1 = 0$.). □

From now on, for simplicity, $T_1 := [0, 0.75] \times [0.75, 1], T_2 := [0.75, 1] \times [0.75, 1], T_3 := [0, 0.75] \times [0.5, 0.75]$, $T_4 := [0.75, 1] \times [0.5, 0.75], T_5 := [0, 0.75] \times [0.25, 0.5], T_6 := [0.75, 1] \times [0.25, 0.5], T_7 := [0, 0.75] \times [0, 0.25]$, $T_8 := [0.75, 1] \times [0, 0.25]$.

As an implication of the remark, there cannot be an equilibrium with (p_M, p_W) in the interior of T_1, T_4 and T_5. If there were such an equilibrium, then $0 < p_M, p_W < 1$. That would imply that there are types who would be willing to Yield. But, the remark shows that in T_1, T_4 and T_5, this is impossible. A similar case can be made to argue that there cannot be an equilibrium in T_8. In short, we can eliminate the values of (p_M, p_W) in T_1, T_4 and T_5 and T_8, since for these values, there is some type of Man who would deviate.

Now, consider values of (p_m, p_w) in T_2, T_3, and T_6, in which $\kappa\lambda > 0$ and $4p_W - 1 > 0$. Here, Yield is psychologically adequate while Assert is materially more profitable. Therefore, the higher is θ^M, the more incentives there are for Man to play Yield. And the perfectly selfish type will prefer Assert. In T_7, the trade-off is the opposite; the higher the type is, the more incentives to assert.

Remark 3. *For values of p_M and p_W in T_2, T_3, and T_6, the types $\theta^M > (<)\theta^* := \frac{4p_W - 1}{[3-4p_W][2p_m + 3p_W - 4p_m p_W - \frac{3}{2}]}$ strictly prefer to play Yield (Assert) whenever θ^* is defined. For values of p_M and p_W in T_7, the types $\theta^M > (<)\theta^*$ strictly prefer to play Assert (Yield) whenever θ^* is defined.*

Proof. In T_2, T_3, and T_6, $\kappa\lambda > 0$. So, Eq. (5) can be written as $\theta^M > (<)\theta^* := \frac{4p_W - 1}{[3-4p_W][2p_m + 3p_W - 4p_m p_W - \frac{3}{2}]}$. A simlar argument holds for T_7, with the inverse inequality. □

Remark 4. *There exist an equilibrium where all types of Man Yield and all types of W assert, and another equilibrium where it is vice versa.*

Proof. W.l.o.g, let $p_M = 0$ and $p_W = 1$. We shall show that no one is willing to deviate. If $p_M = 0$ and $p_W = 1$, then Eq. (5) can be written as:

$$2\theta^M[\frac{3}{2}-2][3-\frac{3}{2}] \geq 3.$$

Clearly, the equation never holds, since the left-hand side must be negative. So, it shows that all types of Man will Assert, and hence $p_M = 0$. No one wants to deviate. Similarly, we can show that no type of women has incentives to deviate from Yield. □

Remark 5. *There exists an equilibrium where all types of both players Yield (i.e. $p_M = 1, p_W = 1$) if and only if $\underline{\theta}^i \geq 6$.*

There exists an equilibrium where all types of both players Assert (i.e. $p_M = 0, p_W = 0$) if and only if $\underline{\theta}^i \geq \frac{2}{9}$.

Proof. ($\Rightarrow$) First of all, note that when $p_M = 1$ and $p_W = 1$, Eq. (5) can be written as $\theta^i \geq 6$. Now suppose that there exists an equilibrium with $p_M = 1$ and $p_W = 1$. Then, all types must be satisfy $\theta^i \geq 6$. And this can only be satisfied if $\underline{\theta}^i \geq 6$. ($\Leftarrow$) Suppose that $\underline{\theta}^i \geq 6$. Then, by having every type Yield, no one would deviate, according to Eq. (5).

The case for (Assert,Assert) can be shown analogously. □

Remark 6. *There exists an equilibrium where $p_M \in [.5,1)$ of Man's types and $p_W \in [.5,1)$ of Woman's types play Yield if and only if there exists a threshold type $\hat{\theta}_i$ for each player such that for all $i, j \in M,W (i \neq j)$,*

$U_i^{\hat{\theta}_i}(Y, p_i, p_j) = U_i^{\hat{\theta}_i}(A, p_i, p_j)$ and $1 - F^i(\hat{\theta}_i) = p_i \in [.5,1)$.

Proof. To prove this, it suffices to note that first of all, since $p_i \in (0,1)$, there are some types who Yield and some types who Assert. Thus, by continuity of the utility function, we know that there must be some indifferent type. Hence, $U_i^{\hat{\theta}_i}(Y, p_i, p_j) = U_i^{\hat{\theta}_i}(A, p_i, p_j)$.

Also, since $p_i \in [.5,1)$ for all i, $\kappa\lambda > 0$. So, Eq. (5) can be written as $\theta_M^* := \frac{4p_W - 1}{[3-4p_W][2p_M+3p_W-4p_M p_W-\frac{3}{2}]}$, and all the higher types will Yield. Therefore, $1 - F^i(\hat{\theta}_i) = p_i \in [.5,1)$. While I showed the remark for ($\Rightarrow$)-direction, the converse can also be shown analogously.

Also, note that the proofs of Remark 6 and 7 are almost identical, so we shall skip the proofs. □

Remark 7. *There exists an equilibrium where $(p_M, p_W) \in (0, \frac{1}{4}]^2$ iff there exists a threshold type $\hat{\theta}_i$ for each player such that for all $i, j \in M,W (i \neq j)$,*

$U_i^{\hat{\theta}_i}(Y, p_i, p_j) = U_i^{\hat{\theta}_i}(A, p_i, p_j)$ and $F^i(\hat{\theta}_i) = p_i \in (0, \frac{1}{4})$.

Remark 8. *There exists an equilibrium where* $(p_M, p_W) \in (0, \frac{1}{4}] \times [.5, .75]$ *iff there exists threshold types* $\hat{\theta}_M$ *and* $\hat{\theta}_W$ *such that* $U_M^{\hat{\theta}_M}(Y, p_M, p_W) = U_i^{\hat{\theta}_i}(A, p_M, p_W)$ *and* $1 - F^M(\hat{\theta_M}) = p_M \in (0, \frac{1}{4})$ $U_W^{\hat{\theta}_W}(Y, p_W, p_M) = U_W^{\hat{\theta}_W}(A, p_W, p_M)$ *and* $F^W(\hat{\theta_W}) = p_W \in (.5, .75)$.

Now that we have discussed how to solve for equilibria under incomplete information, we can discuss the effect of information (or lack thereof) by comparing the acquaintance Society and stranger Society playing BoS. Suppose that individuals in each of the societies are randomly matched to play BoS.[19] The whole population coordinating on (Assert, Yield) or (Yield, Assert) is an equilibrium in both societies.

Observation. *In both societies, the following statements hold:*

(i) There always exists an equilibrium where Man Yields with probability 1 and Woman Assert with probability 1, and another equilibrium where it is vice versa.

(ii) There exists an equilibrium where Man and Woman Yield with probability 1*, if and only if for* $i \in \{M, W\}$ $\underline{\theta}^i \geq 6$.

(iii) There exists an equilibrium where Man and Woman Assert with probability 1*, if and only if for* $i \in \{M, W\}$ $\underline{\theta}^i \geq \frac{2}{9}$.

Proof. We already showed the the observation holds for the stranger society.

Note that the condition for Yield is the same in the acquaintance society.

$$2\theta^M[\frac{3}{2} - 2p_W][2p_M + 3p_W - 4p_m p_W - \frac{3}{2}] \geq (<)4p_W - 1.$$

It is easy to show that, when $p_M = 1$, and $p_W = 0$, no type prefers deviation, which proves (i).

When, $p_M = 1$ and $p_W = 1$, then no type of either player prefers deviation from Yield, if $\underline{\theta}^i \geq 6$. Similarly, in order for $p_M = 1$ and $p_W = 1$ to be achieved in equilibrium, it must be that $\underline{\theta}^i \geq 6$. If it didn't hold, we can find a type less than 6 who would deviate from Yield. Thus, (ii) holds. The proof of (iii) will be almost identical to that of (ii), except that $p_i = 0$. □

Observation. *In the acquaintance society,*

[19] To be precise, we assume that a half of the population is men and the other half is women, and a man is matched with a woman to play BoS.

(i) The maximum probability with which (Y, Y) is played in equilibrium is $[1-F^M(6)][1-F^W(6)]$.

(ii) The maximum probability with which (A, A) is played in equilibrium is $[1-F^M(\frac{2}{9})][1-F^W(\frac{2}{9})]$.

Proof. In order for a pair of players to play (*Y, Y*), it must be the case that $\theta_i \geq 6$, for $i \in M, W$. And this can occur with probability $[1-F^M(\frac{2}{9})][1-F^W(\frac{2}{9})]$. A similar argument holds for (*A, A*). □

It turns out that if (*Y, Y*) can occur with a sufficiently high probability in the stranger society (.625 to be exact) then there exists an equilibrium in the acquaintance society with an even higher frequency of (*Y, Y*).

Proposition. *Suppose that in the stranger society there exists an equilibrium with (Y, Y) frequency higher than* $(.75)^2 = .625$. *Then, there exists an equilibrium in the acquaintance society with a weakly higher (Y, Y) frequency. The result is strict if the maximum frequency in the stranger society is less than 1, and the support of the type distributions have convex supports.*

Proof. In AS, a pair of individuals can play (Y, Y) in equilibrium iff $\theta^M, \theta^W \geq 6$. Therefore, $[1-F^M(6)][1-F^W(6)]$ is the maximum rate of (Y, Y). Now suppose that in the stranger society there exists an equilibrium with (Y, Y) frequency higher than 0.625. The only way this is possible is to have $p_M \geq .75$ and $p_W \geq .75$. If either of the player's Yield probability is less than .75, then the co-player's Yield probability must also be less than .75. That is because equilibria can only occur in $[0,.25]\times[.5,.75]$, $[.5,.75]\times[.5,.75]$, $[0,.25]\times[0,.25]$, and $[.5,.75]\times[0,.25]$. In all of these cases, if $p_M \leq .75$ then $p_M \leq .75$ and vice versa.

So, $p_M \geq .75$ then $p_M \geq .75$. And since this is an equilibrium, there must exist threshold types θ_M^*, θ_W^* such that $1-F^M(\theta_M^*) = p_M$ and $1-F^W(\theta_W^*) = p_W$. Since θ_M^* satisfies the indifference condition,

$$\theta_M^*(p_M, p_W) = \frac{4p_W - 1}{2(\frac{3}{2} - 2p_W)(2p_m + 3p_W - 4p_M p_W - \frac{3}{2})}.$$

In the following lemma, we'll show that $\theta_M^*(x,y)$ is strictly decreasing in $(.75,1)\times(.75,1)$.

Thus, $\theta_M^*(p_M, p_W) \geq \theta_M^*(1,1) = 6$.

$$p_M = 1 - F^M(\theta_M^*) \leq 1 - F^M(6).$$

Similarly, we can show that $p_W \leq 1-F^W(6)$. $\Rightarrow p_M p_W \leq (1-F^M(6))(1-F^W(6))$.

To show that the result is strict, note that if $p_M < 1$ and $p_W < 1$, then $\theta_M^* \in Supp(F^M)$.

And $\theta_M^*(p_M, p_W) > \theta_M^*(1,1) = 6$. So, $p_M = 1 - F^M(\theta_M^*(p_M, p_W)) < 1 - F^M(6)$. Similarly, we can show that $p_W < 1 - F^W(6)$. So, $p_M p_W < (1 - F^M(6))(1 - F^M(6))$.

Now, we shall show that the threshold function is decreasing in p_M and in p_W. The threshold hold function

$$\theta_M^*(x,y) := \frac{4y-1}{2(\frac{3}{2}-y)(2x+3y-4xy-1.5)}.$$

$$\partial_x \theta_M^*(x,y) = \frac{-(2-4y)(4y-1)(3-4y)}{(3-4y)^2(2x+3y-4xy-1.5)^2}.$$

This partial is negative if $p_W \in (\frac{3}{4}, 1]$.

$$\partial_y \theta_M^*(x,y) = \frac{-(3-4x)(4y-1)(3-4y) + 4(3-4x)(y-.5)(3-4y) + 4(4y-1)(y-.5)(3-4x)}{(3-4y)^2(2x+3y-4xy-1.5)^2}.$$

This partial is negative if $p_W \in (\frac{3}{4}, 1]$ and $p_M \in (\frac{3}{4}, 1]$. □

Appendix C

C.1 Proof of Proposition 5

Proof. We only need to discuss s_1 and then the case of s_2 follows symmetrically.

If for all $\theta_1 \in \Theta_1$, $s_1(\theta_1) = D$, then s_1 is a trivial threshold strategy.

If there exists some $\theta_1 \in \Theta_1$ such that $s_1(\theta_1) = C$, then because $U_1^{\theta_1}(C,p) \geq U_1^{\theta_1}(D,p)$, it must be that the utility from taking C is weakly higher than D.

$$p_2 c + \theta_1 \kappa_{12}^{\theta_1}(C,p_2) \cdot \lambda_{121}^{\theta_1}(p_1,p_2) \geq p_2 x + (1-p_2)d + \theta_1 \kappa_{12}^{\theta_1}(D,p_2) \cdot \lambda_{121}^{\theta_1}(p_1,p_2)$$

Arranging the inequality we obtain that

$$\theta_1 \lambda_{121}^{\theta_1}(p_1,p_2)[\kappa_{12}^{\theta_1}(C,p_2) - \kappa_{12}^{\theta_1}(D,p_2)] \geq p_2(x-c) + (1-p_2)d > 0$$

Note that $\kappa_{12}^{\theta_1}(C,p_2) - \kappa_{12}^{\theta_1}(D,p_2) > 0$ as C is always a kind action. So C is a strictly better choice for any type $\theta > \theta_1$, which suggests s_1 must be a threshold strategy. □

C.2 Proof of Proposition 6

Proof. ($\Rightarrow$) Suppose that (s_1, s_2) is a cooperative equilibrium in the sense that there is some positive probability of cooperation by either player. It is easy to verify that no player is willing to cooperate if the co-player never cooperates. So, in equilibrium, the cooperation probability must be positive for both players. I.e. $p_i := \int_{\Theta_i} \mathbb{1}[s_i(\theta_i) = c] dF_i > 0$ for $i = 1, 2$. In fact, choose the pair $p = (p_1, p_2)$ as the pair of the cooperation probability of the players.

And since we already showed that the equilibrium strategy is a threshold function, we just need to make sure that the threshold given by $\theta_i^*(p)$ constitutes the equilibrium strategy.

As shown in Eq. (2), $\theta_i \geq \theta_i^*(p)$ implies that C is optimal (with indifference at $\theta_i = \theta_i^*(p)$) and $\theta_i < \theta_i^*(p)$ implies that D is optimal. So, in fact, the only possible strategy with cooperation probabilities, p, is

$$s_i(\theta_i) = \begin{cases} C & \text{if } \theta_i \geq \theta_i^*(p), \\ D & \text{if } \theta_i < \theta_i^*(p). \end{cases}$$

Now I show that $1 - F_i(\theta_i^*(p)) = p_i$ is satisfied. This can be shown by noting that

$$p_i = \int_{\Theta_i} \mathbb{1}[s_i(\theta_i) = c] dF_i = \int_{\Theta_i} \mathbb{1}[\theta_i \geq \theta_i^*(p)] dF_i = 1 - F_i(\theta_i^*(p)).$$

($\Leftarrow$) Suppose that there is a pair (p_1, p_2) that satisfies the conditions. I show that, then, the strategies,

$$s_i(\theta_i) = \begin{cases} C & \text{if } \theta_i \geq \theta_i^*(p), \\ D & \text{if } \theta_i < \theta_i^*(p). \end{cases}$$

constitute an equilibrium. Since $1 - F_i(\theta_i^*(p)) = p_i$, given the strategies, the cooperation probability of i is p_i. And given this information, for all $\theta_i \geq (<) \theta_i^*(p)$, C (D) is optimal. So, every type is indeed utility-maximizing.

□

C.3 Proof of Proposition 7

Proof. Based on the remark at the end of Section 4, we only need to consider the case in which $p_i > \frac{1}{2}$, $i = 1,2$, and in this area $\theta_i^*(p)$ is decreasing in p_i and p_j according to Eq. (8).

In the acquaintance society, the random draws θ_1 and θ_2 from $F_1(\cdot)$ and $F_2(\cdot)$ are revealed to players. Indeed, the threshold for cooperation equilibrium under complete information is equivalent to $\theta_i^*(1,1)$ by definition. So the two players can form a reciprocity equilibrium (C,C) if and only if $\theta_i \geq \theta_i^*(1,1)$ for $i = 1,2$. That means the corresponding cooperation rate is $[1-F_1(\theta_1^*(1,1))]\cdot[1-F_2(\theta_2^*(1,1))]$.

On the other hand, any SRE (s_1,s_2) in the stranger society with associated probabilities $p = (p_1,p_2)$, such that $p_i \leq 1$ for $i = 1,2$, has a cooperation rate as $[1-F_1(\theta_1^*(p))]\cdot[1-F_2(\theta_2^*(p))]$.

Then, because θ_i^* is decreasing in p, so that $\theta_i^*(p) \geq \theta_i^*(1,1)$ and $[1-F_1(\theta_1^*(1,1))]\cdot[1-F_2(\theta_2^*(1,1))] \geq [1-F_1(\theta_1^*(p))]\cdot[1-F_2(\theta_2^*(p))]$. □