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 reciprocity model and explore its implications for reciprocal behavior. We extend the reciprocity model in extensive form games (Dufwenberg and Kirchsteiger, 2004), develop the Extended Sequential Reciprocity Equilibrium (ESRE), and prove its existence. We use this general framework to study many well-known games, by comparing the theoretical predictions in complete and incomplete information games. We find that, in prisoners' dilemma, players are more likely to cooperate with each other when they have information about the reciprocal motivations of their opponents, given the benefit of defection is not too large.

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 and it is tempting to relax the assumption of common knowledge about the intensity of these motivations. As Attanasi, Battigalli and Manzoni (2016) have argued, it is implausible that the subjects who are randomly drawn from a population to participate in an experiment would have sufficient information to know others' other-regarding preferences. Furthermore, different experiments have suggested large heterogeneity in social motives among individuals. For example, Dohmen, Falk, Huffman and Sunde (2008) report, based

* Corresponding author

on a survey, that people show a high degree of heterogeneity in trust and reciprocity among individuals. Hennig-Schmidt, Sadrieh and Rockenbach (2010) show that there is large heterogeneity among employee's effort level after the employer offered them a bonus. Bellemare, Sebald and Suetens (2018) have also documented evidence that, in the dictator game, the dictators exhibit heterogeneous guilt sensitivities. These laboratory experiments suggest that the assumption of commonly known psychological motivations may not be innocuous in many situations.

There have been papers that discuss the implications of assuming incomplete information in psychological games. Battigalli and Dufwenberg (2009) lay out a general foundation for psychological games where they stress the necessity of extending 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. We attempt to make contributions in this direction, as well. In particular, we focus on a reciprocity game in which players are uncertain about the intensity of each other's psychological motivations.

We extend the reciprocity model in extensive form games (Dufwenberg and Kirchsteiger,2004; henceforth DK) to incorporate incomplete information about sensitivity parameters, which are the degrees to which players care about social motives. Previous reciprocity models incorporate into standard game theory people's natural tendency to reward kind people and punish mean people. But they maintain the assumption of complete information about reciprocal motivations (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010). Therefore, our model complements to this strand of literature by facilitating the analysis of reciprocal behavior subject to the uncertainty about others.

As Sebald (2010) points out, chance moves in a reciprocity model could alter the way players attribute responsibilities. In this paper, the uncertainty of sensitivity parameters causes players' psychological motivations to be moderated. Because when one player lacks information about others, the consequence of his action depends on what other players would do based on their information. 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*. As a result, his kindness would be evaluated in terms of how his action affects the *expectation* of other players' well-being.

However, the direction of the impact of uncertainty on equilibrium predictions is ambiguous. In many

games, reciprocal equilibrium outcomes are robust to a relatively small probability that players might be selfish. As will be shown in Section 5.3, in an ultimatum game, the proposer would make a positive offer as high as that to a reciprocal responder, despite the fact that the responder has a chance to be selfish and would like to accept the lowest offer. The similar patterns are also found in monopoly pricing and public goods games (Sections 5.2 and 5.5). On the contrary, in other games, the reciprocal equilibrium outcome can be sensitive to the introduction of uncertainty. For instance, in prisoners' dilemma, cooperation can be rather hard to maintain once both parties realize that their opponents could possibly be selfish (Sections 3 and 4).

The class of extensive form games of interest are multi-stage games with observed actions and independent types. In standard game theory, Fudenberg and Tirole (1991) have studied this class of games and define a strong version of Perfect Bayesian Equilibrium. We inherit their restrictions on the belief system, requiring the equilibrium assessment to be *reasonable*. It means that the type distributions of each player are updated independently and any deviation of a player should not signal information that he does not possess. In addition, in the equilibrium notion of our reciprocity model, the Extended Sequential Reciprocity Equilibrium (ESRE), we require the condition of *sequential rationality* that each player should maximize his utility in each continuation game given the specified belief.

Note that Sebald (2010) extends DK by chance moves. However, he only introduces procedural randomize options by using which players can avoid being held responsible for realized outcomes and thus mitigate others' reciprocal motivations. Unlike this paper, he maintains the assumption of complete information about sensitivity parameters. Bierbrauer and Netzer (2016) also contain uncertainty in reciprocity models in the context of mechanism design. In their setup, players have private information that is relevant for material well-being, yet the Revelation Principle does not hold in the presence of psychological motivations. They examine the extent to which the implementable social choice functions are robust to the presence of the psychological motivations. In the construction of certain mechanisms, they exploit the feature of a reciprocity model that agents' reciprocal incentives are effectively influenced by the feasible alternatives.

The outline of our paper is the following. In Section 2, we lay out the reciprocity model in extensive form games. In Section 3, we exemplify by studying prisoners' dilemma and illustrate the implications of uncertainty on reciprocal behavior. In Section 4, we compare the equilibrium outcomes between two "societies" with or without information about members' psychological motivations. In Section 5, we apply

the theory to a series of examples and see how it may change the results from reciprocity models with complete information. In Section 6, we conclude.

2. The Model

We build a reciprocity model in extensive form games with uncertainty about players' sensitivity parameters. The baseline model is DK, who extend the original reciprocity model from simultaneous games Rabin (1993) to extensive form games under the assumption of complete information. The key difference between our framework and DK's is that now players need to update their beliefs about types before making decisions. Since players' actions may vary across types, the change in beliefs about types would influence players' reciprocal motivations and the play of game.

Suppose there are I players. The type space for each player i is Θ_i, and the product of all type spaces is $\Theta = \prod_i \Theta_i$. Note that θ_i is a vector $(\theta_{i1}, \ldots, \theta_{i,i-1}, \theta_{i,i+1}, \ldots, \theta_{iI})$ that consists of $I-1$ elements, where θ_{ij} ($j \neq i$) represents i's sensitivity parameter with respect to j.

At the beginning of the game, nature moves and randomly selects a type θ_i for each player i from a 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 of each player are independent.

The game proceeds for T stages. At stage t, each player makes a decision from a finite set $D_{i,t}$ simultaneously. For simplicity, we assume that all types of one player share the same feasible choice set. At the end of each stage, all the decisions made are observed by all players. Let $D_i = \bigcup_{t=1}^{T} D_{i,t}$ be the set of all feasible actions for player i. A history h^t contains the decisions of all players up until stage t, which belongs to the set H^t. Let $H = \bigcup_{t=1}^{T} H^t$ be the set of all histories and let h_0 be the initial node.

A behavior strategy a_i for player i specifies the (mixed) action each type would take at each stage. That is, a_i specifies a map from $\Theta_i \times H^{t-1}$ to $\Delta(D_{i,t})$ for each t. We denote by A_i the set of all behavior strategies for i and by A the set of all strategy profiles. Since player i has information about his own type, we also need the notation for the behavior strategy associated with a single type. The behavior strategy for a single type θ_i is a mapping $s_i : H \longrightarrow \Delta(D_i)$, so that $s_i(h) = a_i(\theta_i, h)$. We denote by S_i the set of all strategies of single types and S the set of all profiles of such strategies.

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 a little by using $\pi(s_i, a_{-i}, \mu_{-i})$ to denote i's expected payoff when he plays a strategy s_i after he knows his type and holds a belief μ_{-i} about others' types. Since types do not enter the material payoff functions, we can also define players' expected material payoffs in complete information games as functions $\bar{\pi}_i : S \longrightarrow \mathbb{R}$.

At each stage of the game, the players will update beliefs about types based on observed history and initial strategy profile. Following Fudenberg and Tirole (1991), in such a multi-stage game with observed actions, we can assume that they have common knowledge about the beliefs 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 satisfies the property of *consistency* as defined in Kreps and Wilson (1982).

On the other hand, we also ask players to revise beliefs about strategies as game unravels. The discussion of this necessity is contained in DK. Specifically, we ask different types to revise beliefs about strategies in the same way. Suppose initially players hold a belief that a_i will be played by i, then at history $h \in H$, $a_i(h) \in A_i$ is the revised belief about a_i. We follow DK by letting $a_i(h)$ be the same as a_i except for the histories that define h. The interpretation is that in all predecessors of h, players are *believed* to make choices that can lead to h with probability 1 conditional on that h has been reached. For a certain type θ_i, the revised belief about a behavior strategy, $s_i(h)$, is similarly defined, i.e., $s_i(h) = a_i(h)(\theta_i)$.

Then, we introduce players' first- and second-order beliefs about strategies. Player i's belief about j's strategy is contained in $B_{ij}(= A_j)$, and player i's belief about j's belief about k's strategy is contained in $C_{ijk}(= B_{jk})$. When players update their beliefs about strategies at history h, they update their first- and second- order beliefs in parallel; thus, $B_{ij}(h) = A_j(h)$ and $C_{ijk}(h) = B_{jk}(h)$.

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 he can choose from 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* associated with a single type in the same way as in complete information games (DK, 2004; 2019). The idea is that, for any type of a player, an efficient strategy should not be outperformed by another strategy in all histories, regardless of any strategies

played by other players.

The set of efficient strategies associated with a single type for player i is defined as below:

$$E_i = \{s_i \in S_i |\ \nexists s_i' \text{ s.t. for all } h \text{ and } (s_j)_{j\neq i}, \bar{\pi}_i(s_i'(h), (s_j(h))_{j\neq i}) \geq \bar{\pi}_i(s_i(h), (s_j(h))_{j\neq i}), \text{ with at least one strict inequality}\}$$

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

In any continuation game, each player i updates his belief about other players' types $\gamma_{-i} \in \prod_{j\neq i}\Delta(\Theta_j)$ and belief about other players' strategies $(b_{ij})_{j\neq i} \in \prod_{j\neq i} B_{ij}$. 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 player i's view, a fair amount of payoff player j deserves is modeled as the average of the maximum and minimum of the range.

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

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 would think his *kindness* to j at history h as being j's expected payoff, which depends on i's behavior strategy, relative to the equitable payoff to j. 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

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

Conversely, in player i's view, the equitable payoff to him from j's play, conditional on his second-order

belief $(c_{ijk})_{k\neq j}$ and belief about type distribution γ, is given by

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

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,

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

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 $U_i^{\theta_i}: S_i\times\prod_{j\neq i}\left(Bij\times\prod_{k\neq j}C_{ijk}\right)\times\Delta(\Theta)\to\mathbb{R}$, which can be separated into material and psychological payoffs. Specifically,

$$\begin{aligned}&U_i^{\theta_i}(s_i(h),(b_{ij}(h),(c_{ijk}(h))_{k\neq j})_{j\neq i},\mu(h))\\ =&\pi_i(s_i(h),(b_{ij}(h))_{j\neq i}),\mu(h))\\ &+\sum_{j\neq i}\theta_{ij}\cdot\kappa_{ij}\left(s_i(h),(b_{ij}(h))_{j\neq i},\mu_{-i}(h)\right)\lambda_{iji}(b_{ij}(h),(c_{ijk}(h))_{k\neq j},\mu(h))\end{aligned}$$

In the equilibrium analysis, we treat each player as being rational "agents" 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 each agent maximizes his "local" utility.

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

Finally, we define the equilibrium notion as below.

Definition 1. (a^*,μ^*) *is an Extended Sequential Reciprocity Equilibrium (ESRE) if:*

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

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

(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* reasonable *in the sense of Fudenberg and Tirole (1991).*

According to Definition 1, 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, therefore the first- and second-order beliefs coincide with the equilibrium strategy profile. Condition (2) requires that the belief updating obeys Bayes' rule at each information set that is reached with positive probability and that in "zero-probability" events players would hold beliefs that satisfy the "no-signaling-what-you-don't-know" condition proposed by Fudenberg and Tirole (1991).

Since condition (2) is implied by *consistency* in the sense of Kreps and Wilson (1982), it suffices to prove existence of an assessment that satisfies sequential rationality and consistency. The proof is a combination of that of sequential reciprocity equilibrium in DK and that of trembling-hand equilibrium in Selten (1975). We first prove the existence of equilibrium in perturbed games where players are restricted to play completely mixed strategies and the entire belief system is pinned down by Bayes rule. This proof resembles that in DK. Then we exploit the upper semi-continuity of the sequence of equilibria in perturbed games. It can be shown that there exists such a sequence of equilibria in perturbed games that converges to an assessment that satisfies sequential rationality and consistency.

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

3. Prisoners' Dilemma with Private Sensitivities

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

In Prisoners' Dilemma as well as Battle of the Sexes in Section 5.1, we assume continuous type spaces for the simplicity of exposition, i.e., $\Theta_i = [\underline{\theta}_i, \bar{\theta}_i] \subset \mathbb{R}_+$.[2] 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 $\theta_i < \theta_i^*$, $s_i(\theta_i) = D$.[3] In Proposition 1, 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. Plus 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

[2] In the general model, we assume finite type spaces to facilitate the existence proof of an equilibrium. In PD and BoS, however, an equilibrium always exists. In addition, the deviation from finiteness does not change the main insights of the examples.

[3] 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.

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 1. *Suppose a pair of strategies (s_1, s_2) is an ESRE. 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 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 person and make 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 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 ESRE 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" ESRE 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.(4) the interim utilities of type θ_i from taking C and D can

be written as:

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

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.

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

where $p_i \neq \frac{1}{2}$ and $(x-d)+(c+d-x)p_i \neq 0$ for $i = 1,2$.[4] Now that Eq.(2) 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 1, the characterization of the ESRE in prisoners' dilemma is as follows.

Proposition 2. *(s_1, s_2) is a cooperative ESRE 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.
$$1 - F_i(\theta_i^*(p)) = p_i. \tag{3}$$

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

[4] 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.

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. Stranger vs. Acquaintance Societies

In this section and Section 5, we study the equilibrium outcomes in *stranger* and *acquaintance* societies. Let us explain the meanings of these two types of societies. For any given game form, players have a common prior about each player's type distribution. Then, their types are randomly drawn according to the distributions and privately informed to each player. In the acquaintance society, players start the game with the knowledge of each other's types. In the stranger society, players start the game without the knowledge, yet they may update their beliefs as game proceeds.

4.1. The Condition under which Information Reinforces Cooperation

This section examines the effect of information on cooperation in PD by comparing the mutual cooperation rates, the probabilities of achieving the socially optimal outcome (C,C), in the stranger and acquaintance societies. 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. 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. (2) to disentangle the material and reciprocal effects. As in Eq. (4), 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. 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.

$$\theta_i^*(p) = \frac{1}{2} \cdot \frac{\pi_i(D, p_j) - \pi_i(C, p_j)}{\kappa_{ij}^{\theta_i}(C, p_j) \cdot \lambda_{iji}^{\theta_i}(p_i, p_j)} \tag{4}$$

From Eq. (4), 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 an 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 cooperation becomes 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 3. *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. 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.

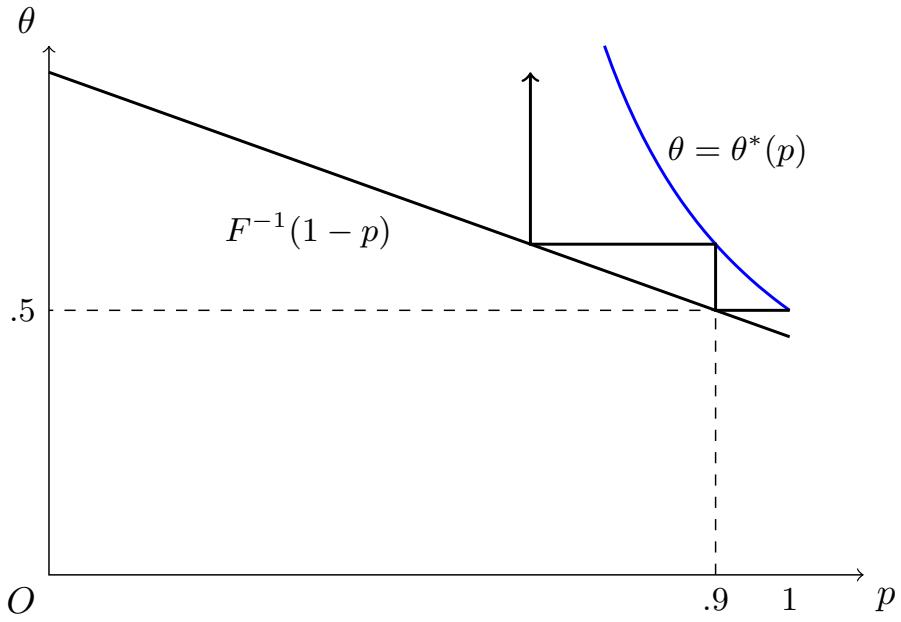


Figure 1: The Collapse of Cooperation

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 by 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 of being 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 1 illustrates 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 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.

4.3. When Information Can Hinder Cooperation

In Section 4.1, 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 society achieves 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 PD game form 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 the stranger society, the equilibrium prediction is a player is willing to cooperate if and only if her type is greater than or equal to 0.372. We can compute that stranger society achieves mutual cooperation with probability $(.82)^2 \approx .67$, and acquaintance society achieves mutual cooperation with probability $(.8)^2 = .64$ (Figure 2). 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.

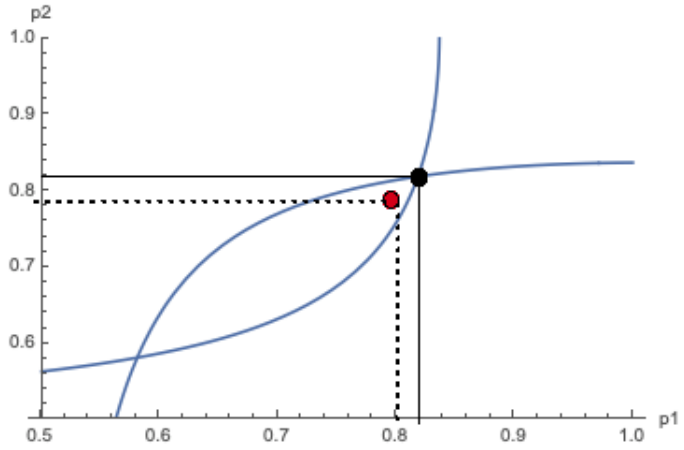


Figure 2: Equilibrium Outcome p_1 and p_2 in Acquaintance Society (Red) and Stranger Society (Black)

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 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. Applications

In this section, we apply the theoretical model to a series of games and illustrate the implications of uncertainty for reciprocal behavior. We compare equilibrium outcomes in stranger and acquaintance societies and find inspiring results under parametric conditions.

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

Figure 3: Battle of the Sexes

5.1. Battle of the Sexes

In this section, we study the Battle of the Sexes (BoS) in Figure 3 and discuss the implication of reciprocal uncertainty.[5] Call the player roles Man (M) and Woman (W). Suppose that players' reciprocal types, θ^M and θ^W are random variables with cumulative distribution functions, F^M and F^W respectively, each with the interval support, $\text{supp}(F^i) = [\underline{\theta}^i, \bar{\theta}^i]$, for $i \in \{Man, Woman\}$. We shall again suppose that a large population of agents are randomly paired to play the BoS, and we perform comparative statics of the two environments: the acquaintance and stranger societies. We focus on comparing the equilibrium behavior and we defer the calculation to Appendix B. In the following observation, we compare the equilibrium condition where all the types of each player plays the same action.

Observation 1. *In both societies, the following statements hold:*

(i) There always exists 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 for* $i \in \{Man, Woman\}$ $\underline{\theta}^i \geq 6$.

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

Observation 1-(i) is intuitive; regardless of their types, players are materially maximizing, and also they are being kind to the kind co-player. Also, in order for both players to Yield with probability 1, which is materially worse off, they must be sufficiently reciprocal to have incentives to punish each other for being unkind. And (Yield,Yield) requires a stronger reciprocity sensitivity than (Assert, Assert), as the players forgo more material payoff to punish the co-player by a smaller amount. The implication of this observation

[5] Note that while BoS 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.

is that if we compare the equilibria where all the types play the same actions, the availability of information makes no difference. In particular, it implies that if we compare the maximum probability of the "efficient outcomes," (i.e., (Y,A) or (A,Y)) information plays no role.

For a more interesting case, we compare the maximum potential for the *inefficient* outcomes, (Y,Y), and (A,A) in the two informational settings. Can we make a statement about which society may be *better* at staving off inefficient outcomes? First, the following observation describes the maximum probability of (Y, Y) and for (A, A) in equilibrium in the acquaintance society.

Observation 2. *In the acquaintance society,*

(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 $[1-F^M(\frac{2}{9})][1-F^W(\frac{2}{9})]$.

Proposition 4 compares the maximum probability of (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 is achieving a high rate of (Y,Y) in equilibrium without information about each other, it may be *worse* to provide information about each other. The rough intuition is as follows. Let p_i be the probability of i's Yield. Note that first of all, the types of Man who are willing to play Y, knowing that the co-player plays Y (i.e. $p_W = 1$) must be strongly reciprocal, as they are giving up 3 to punish 1. And now, suppose that the player is now slightly unsure: the co-player plays Y with $p_W < 1$. Then, this has two effects on utility. First, Y becomes slightly more profitable than when $p_W = 1$. Second, since Woman is less unkind, and one's incentives for punishment decreases. As a result A becomes psychologically more attractive. The types who play Y in the acquaintance society are reciprocal enough to care about the second, psychological effect more than the first, material effect. The upshot is that there are some types who play Y in the acquaintance society and play A in the stranger society.

One may suspect that an analogous result may hold for (A,A). It turns out that is not the case. Consider the equilibrium with a highest rate of (A,A). Man does not need to have a high type to play A against A, since the material gain from deviation to Y is 1, but he can punish by 3. So, it is psychologically easy

to justify A against A (i.e. $p_W = 0$). Now, suppose that Man is slightly unsure Woman is playing A (i.e. $p_W > 0$). Two effects are also at play. First, A becomes materially more profitable. Second, Y becomes psychologically more attractive, as the willingness to punish decreases. And, some of the types who are playing A are low enough to care about the effect of "increasing material benefit" of A, than the effect of "diminishing willingness to punish." So, the types who played A in the acquaintance society will stick with A in the stranger society. And, there are types who are selfish enough to prefer Y against A, but when $p_W > 0$, switch to A.

5.2. Monopoly Pricing

One context in which reciprocity has been first discussed is *monopoly pricing*. Rabin (1993) has shown that when consumers are motivated by reciprocity,[6] the consumer will refuse to buy from the monopolist if the price is higher than what they deem fair. And the profit-maximizing monopolist must, therefore, set the price lower than the monopoly pricing level. We will apply our framework to explore how the monopoly pricing would change when the firm is unsure of the consumer's reciprocal motivation.

Consider a profit-maximizing monopolist (M) who produces a good which costs c per unit, and a 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 $\theta > 0$. $Prob(\theta_C = 0) = \eta \in (0,1)$. 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. In addition, we will focus on the cutoff strategies for the type θ. In other words, θ 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 a pair (p, r).[7]

It turns out that there is a continuum of equilibria where the equilibrium 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 3. *In any equilibrium, the reservation price r for the reciprocal type of C lies in* $[\underline{p}, \bar{p}]$.

[6] Rabin (1993) calls it fairness.

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

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 take advantage of the selfish type and let the reciprocal type *refuse*. If the probability of the selfish type η is sufficiently large, M will charge $p = v$ in equilibrium, and the reciprocal type will *refuse* to buy. If η is sufficiently small, M will charge the reservation price, r and both types will buy at that price.

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 reasonable 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 4. *If the probability of the selfish type, $\eta \leq \frac{1}{2}$, there is an equilibrium where both types choose 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 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 5. *If $\eta \leq \frac{1}{2}$,*

- *When the highest-reservation-price equilibria are compared, M's equilibrium price in Stranger Society is identical to the price for the reciprocal consumer in Acquaintance Society.*
- *When the lowest-reservation-price equilibria are compared, M's equilibrium price in Stranger Society is strictly higher than the price for the reciprocal consumer in Acquaintance Society.*

One takeaway from (i) is that if more than a half of the population is reciprocal, the selfish individual will benefit in Stranger Society, as they will face a price lower than that in Acquaintance Society. And the reciprocal individuals are not worse off as they face the same price as in Acquaintance Society. In short, consumers are weakly better off in Stranger Society if $\eta \leq \frac{1}{2}$, when the highest-reservation-price equilibria are compared. 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, and in this case, the reciprocal type will face a lower price in Acquaintance Society.

The intuition for (ii) the result is as follows. If we compute the reservation price that will be accepted by the consumer, then for any price higher than that, the consumer must not buy. Then, we need to compute the lowest possible price that will be *refused* by the consumer, and this depends on how unkind C deems M to be (because C would refuse only when M is deemed unkind). If C deems M to be more unkind, C is more willing to *refuse*, so M needs to offer a lower price to be accepted. And recall that if M believes C is about to *refuse* at p, then by offering p M is more unkind in Acquaintance Society than in Stranger Society, and so C is more lenient on M in Stranger Society. Thus, M can get away with charging a higher price in Stranger Society. As a result, the reciprocal type will face a higher price in Stranger Society than what is predicted by the complete information model.

5.3. Ultimatum Game

In a typical ultimatum game, there are a proposer (P) and a responder (R). We 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.[8]

Suppose R's sensitivity parameter is θ_R and it is public information, and 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.

[8]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 6. *In P's favorite equilibrium, he would offer R an amount equal to*

$$\bar{s}(\theta_R) = \frac{1}{4}\left[(3+\frac{2}{\theta_R}) - \sqrt{(3+\frac{2}{\theta_R})^2 - 8}\right] \tag{5}$$

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

Suppose there are two types, a selfish type with $\theta_R = 0$ and a reciprocal type with $\theta_R > 0$. Let the prior probability of the selfish type be $(1-\mu) \in (0,1)$. In equilibrium, the selfish type plays a strategy that prescribes acceptance in response to any offer. So we only need to discuss the choice of the unselfish type. Let $\bar{s}(\theta_R)$ be the cutoff level for the unselfish type as in the complete information game with $\theta_R > 0$.

We explore two possibilities:

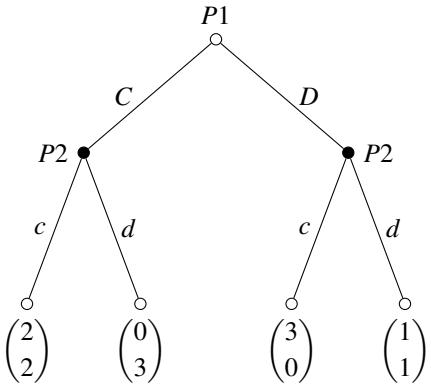
1. $x = 0$. The selfish type would accept and the unselfish type would reject. P's expected payoff is $(1-\mu)$.
2. $x = \bar{s}(\theta_R)$. If it was ever an equilibrium offer, R must accept it; otherwise, P can deviate to $x = 0$. Then, given that P expects that both selfish and unselfish types will accept it, R's belief about P's kindness to her is the same as in the complete information case when a deal is reached. In addition, the kindness of R to P is also the same. That means the cutoff level for R is the same as $\bar{s}(\theta_R)$. Whenever P wants to induce acceptance by R, he offers exactly $\bar{s}(\theta_R)$. In such an equilibrium, his payoff is $\bar{s}(\theta_R)$.

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 selfish type would accept. It is obviously more profitable for P to deviate to $x = 0$.

Proposition 7. *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 $\bar{s}(\theta_R) < \frac{1}{2}$, when $\mu \geq \frac{1}{2}$, P will offer $\bar{s}(\theta_R)$.*

The economic intuition is clear. When the majority of population is reciprocal, the proposer should increase the offer to the level as in the complete information game with respect to the reciprocal type.

5.4. Sequential Prisoners' Dilemma



In a sequential PD game, the second player (he) can condition his decision on the first player (she)'s choice. 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, the unselfish type of P2 may want to choose c when she is strongly motivated by reciprocity.

To support the choice of c, it needs to be satisfied that: $\theta_2 \geq \frac{1}{2-p_2}$. Note that 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 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. So she is more grateful for P1's generosity when he truly chooses C.

Observation 5. *Given that P1 has chosen C, it is optimal for P2 to choose c if* $\theta_2 \geq \frac{1}{2-p_2}$.

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

Observation 6. *Case 1: When P2 is more likely to be reciprocal ($p_2 \geq \frac{1}{2}$), mutual cooperation is sustained as an equilibrium outcome if $\theta_2 \geq \frac{1}{2-p_2}$.*

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 P2 is more likely to be selfish ($p_2 < \frac{1}{2}$) and P1 is more likely to be reciprocal ($p_1 \geq \frac{1}{2}$), then mutual cooperation is possible when the unselfish types of 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 7. *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 6. 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; plus, P2 is overall kind, and he wants to return the favor by taking C.

5.5. Public Goods Game

We revisit the public goods game studied by Dufwenberg, Gachter and Hennig-Schmidt (2011) and introduce uncertainty to see what difference it makes. In this game, there are 3 players, each of whom begins with 20 tokens as the endowment. Each player i can contribute s_i ($s_i \in [0,20]$) tokens to producing

public goods G with the production function $G = \frac{3}{2}(s_1 + s_2 + s_3)$. All the public goods will be equally shared by the players. Therefore, each player will receive $\frac{1}{3}G = \frac{1}{2}(s_1 + s_2 + s_3)$. The payoff to player i equals the remaining tokens, $20 - s_i$, and public goods he receives, $\frac{1}{3}G$.

Let b_{ij}^1 and b_{ik}^1 denote player i's first-order beliefs about players j and k's contributions. Then player i would prefer to contribute all his tokens when the total contributions from other players are so high such that

$$\theta_i \geq \frac{2}{b_{ij}^1 + b_{ik}^1 - 20} \tag{6}$$

Suppose that all players break ties by contributing to the public good. Then there exist a symmetric equilibrium where all players contribute only when the sensitivity parameter of each player satisfies that $\theta_i \geq \frac{1}{10}$.

Let us introduce uncertainty by assuming that the type of each player θ_i could take on two values $\underline{\theta}_i$ and $\bar{\theta}_i$ ($\underline{\theta}_i < \bar{\theta}_i$) and the prior probability of $\bar{\theta}_i$ equals p. If a player is selfish, then he would give out nothing, which is an additional risk for any player who considers contribution.

Observation 8. *The incentive constraint for the type θ_i of player i to contribute his tokens is that*

$$\theta_i \geq \frac{2}{\mathbb{E}(b_{ij}^1) + \mathbb{E}(b_{ik}^1) - 20} \tag{7}$$

Notice that Ineq. (6) is similar to Ineq. (7) except that it requires the *expectation* of others' contributions to reach a certain level.

Because a low value of p would lower the expectation of players' about each other's contributions, it can further hinder the chance of reaching the socially optimal outcome, i.e., everybody contributes all of their tokens to the public goods. Under this parametric setting, there could be no contribution, full contribution, and partial contribution in equilibrium.

Observation 9. *The symmetric equilibria may take three forms:*

- *1. If $p < \frac{1}{2}$, then in any equilibrium no player contributes to the public good.*
- *2. If $p \geq \frac{1}{2}$, $\underline{\theta}_i \geq \frac{1}{10}$, and $\bar{\theta}_i \geq \frac{1}{10}$, then there is an equilibrium where all players contribute $s_i = 20$.*

- *3. If $p \geq \frac{1}{2}$ and $\underline{\theta}_i < \frac{1}{10}$, then there is an equilibrium where the type $\bar{\theta}_i$ of all players would like to contribute 20 tokens when*

$$\bar{\theta}_i \geq \frac{1}{10(2p-1)} \tag{8}$$

Note that in the third case, the lower bound $\frac{1}{10(2p-1)}$ is always higher than that in the complete information, i.e. $\frac{1}{10}$. That means in the presence of possible selfish types, players are less willing to cooperate even if they are themselves strongly motivated by reciprocity.

If we focus on the equilibria with the highest contribution levels, then we have following results.

Observation 10. *If $\underline{\theta}_i, \bar{\theta}_i < \frac{1}{10(2p-1)}$, or $\underline{\theta}_i, \bar{\theta}_i \geq \frac{1}{10}$, then the acquaintance society has a weakly higher contribution level in expectation; if $\underline{\theta}_i < \frac{1}{10} < \frac{1}{10(2p-1)} \leq \bar{\theta}_i$, then the stranger society has a higher contribution level in expectation.*

To understand the first part of this result, note that when $\underline{\theta}_i, \bar{\theta}_i \leq \frac{1}{10(2p-1)}$, the players will not offer any contribution in the stranger society, whereas it is possible that the high types of all players can coordinate in positive contributions in the acquaintance society. When $\underline{\theta}_i, \bar{\theta}_i \geq \frac{1}{10}$, both types of players would like to offer positive contributions in the acquaintance society, so the expected contribution level is 20.

In the third case, when $\underline{\theta}_i < \frac{1}{10} < \frac{1}{10(2p-1)} \leq \bar{\theta}_i$, the stranger society outperforms the acquaintance society. Because in the stranger society, the high type of each player would like to contribute, which results in the expected contribution level as $60p$. In the acquaintance society, however, the high type of each player wants to contribute only when all other two players are of high types. So the expected contribution level is $60p^3$.

5.6. Investment Game with Punishment

The purpose of this game is to illustrate how the type updating feature of our model influences equilibrium behavior, and in-so-doing, provides a rationale for our modeling choice of how the updating is performed. This game can be thought of as a form of the "investment game" with punishment. Player 1 (she) is an investor and Player 2 (he) is an entrepreneur. The investor can invest or not. Given the investment, the entrepreneur can put efforts and share the profits so they both earn \$2, or slack off and keep \$3

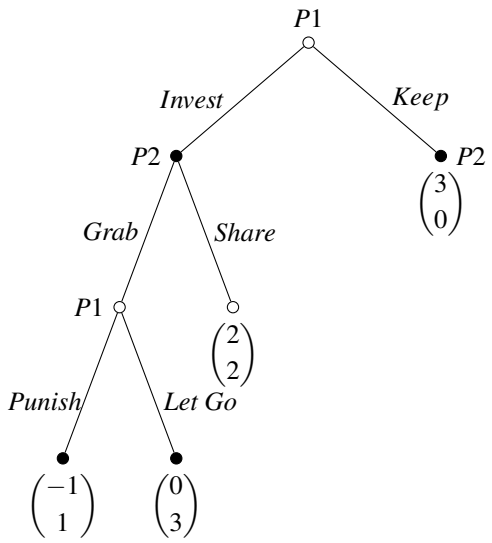


Figure 4: Investment Game with Punishment

for himself. If the entrepreneur slacks off, the investor can pay \$1 to punish by \$2. This results in the game form given in Figure 4.

Suppose that P2 is selfish ($\theta_2 = 0$), and P1 could be either selfish ($\theta_1 = 0$) with probability .6 or reciprocal ($\theta_1 = \tilde{\theta}_1 > 1$) with probability .4. The assumption, $\tilde{\theta}_1 > 1$ is to ensure that the reciprocal type will punish when given the option. It can be checked that if $\tilde{\theta}_1 > 1$, then the reciprocal type of P1 would certainly punish P2, if P2 grabs.

First off, note that the selfish type of P1 will always *keep* at the root, since *invest* will necessarily give him a lower payoff. So, if P2 is called to move with any positive probability, it is reasonable for P2 to presume that he is facing the reciprocal type. What shall P2 do in this case? Since he is selfish, he doesn't care about the kindness of P1. He chooses to *grab* if P1 will *let go* with probability greater than .5. Otherwise, he would *share*. But, since P2 believes that P1 is reciprocal, P2 also believes that P1 will *punish* with probability 1 if he *grabs*. So, P2 has to *Share*. This is supported in equilibrium: P1 plays (invest, punish) if reciprocal, and (keep, let go) if selfish. P2 *shares*.

So, in this example, the punishment option can serve as a credible threat only when P1's *invest* reveals P1's type. And in this case, the incomplete information game would result in the same predictions as the

complete information game.

This example illustrates why a psychologically plausible model of reciprocity must keep track of both "action revising" (*à la* DK) and "type updating."

Suppose that players do not *revise* beliefs about actions. Then, the reciprocal type of P1 may not *punish* in equilibrium, even if P1 observed that P2 *grabbed*. After observing P1's *grab*, P2 does not necessarily think that P2's Grab was intentional. This is psychologically implausible and does not capture the idea of reciprocity well. But, the action-revision feature *per se* is not enough, without type updating. Say that P2 does not perform type updating, and given P1's *investment* believes that all types of P1 are investing. Since P2 believes that with probability .6, P1 is selfish P1 will *let go* with probability .6. P2 finds it maximizing to *grab*. In this case, investment will never be supported in equilibrium. However, as we elaborated, rationality implies that P2 infers that he is facing the reciprocal type, since the selfish type would have chosen *Keep* instead. So, a psychologically sound model with rationality should predict that P2 must share. It leads to our model of reciprocity that keeps track of both action-revising and type-updating.

References

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

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.

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.

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

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.

Selten, R., 1975. Reexamination of the perfectness concepts for equilibrium points in extensive games. International Journal of Game Theory 4.1, 25–55.

Appendix A

A.1 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 $\mu(a)$.

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, 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 $\mu(a_n) \to \mu^*$ and $a_n \to a^*$.

At last, we show that (a^*,μ^*) is an equilibrium assessment. 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 $a_i^*(\theta_i)(h)$ to $s_i \in A_i(\theta_i,h,a^*)$ and

$$U_i^{\theta_i}(s_i,(b_{ij}^*(h),(c_{ijk}^*(h))_{k\neq j})_{j\neq i},\mu^*(h)) > U_i^{\theta_i}(a_i^*(\theta_i)(h),(b_{ij}^*(h),(c_{ijk}^*(h))_{k\neq j})_{j\neq i},\mu^*(h))$$

where $b_{ij}^*(h) = a_j^*(h)$ and $c_{ijk}^*(h) = a_k^*(h)$, for all i,j,k.

Then we construct a sequence of strategies s_i^n such that s^n coincides with $a^n(\theta_i)(h)$ except for h, and $s_i^n(h) = s_i(h)$. Because $U_i^{\theta_i}$ is continuous, when n is large enough, $U_i^{\theta_i}(s_i^n,(b_{ij}^n(h),(c_{ijk}^n(h))_{k\neq j})_{j\neq i},\mu^n(h))$ approximates $U_i^{\theta_i}(s_i,(b_{ij}^*(h),(c_{ijk}^*(h))_{k\neq j})_{j\neq i},\mu^*(h))$ and $U_i^{\theta_i}(a_i^n(\theta_i,h),(b_{ij}^n(h),(c_{ijk}^n(h))_{k\neq j})_{j\neq i},\mu^n(h))$ approximates $U_i^{\theta_i}(a_i^*(\theta_i)(h),(b_{ij}^*(h),(c_{ijk}^*(h))_{k\neq j})_{j\neq i},\mu^*(h))$. That means there is n such that

$$U_i^{\theta_i}(s_i^n,(b_{ij}^n(h),(c_{ijk}^n(h))_{k\neq j})_{j\neq i},\mu^n(h)) > U_i^{\theta_i}(a_i^n(\theta_i)(h),(b_{ij}^n(h),(c_{ijk}^n(h))_{k\neq j})_{j\neq i},\mu^n(h))$$

where $b_{ij}^n(h) = a_j^n(h)$ and $c_{ijk}^n(h) = a_k^n(h)$, for all i,j,k. That means, 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 $\Gamma(\varepsilon_n)$. □

A.2 Proof of Proposition 1

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. □

A.3 Proof of Proposition 2

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.

□

A.3 Proof of Proposition 3

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 ESRE (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 $[1-F_1(\theta_1^*(1,1))]\cdot[1-F_2(\theta_2^*(1,1))] \geq [1-F_1(\theta_1^*(p))]\cdot[1-F_2(\theta_2^*(p))]$. □