Information Disclosure and Cooperation in a Finitely-repeated Dilemma: Experimental Evidence

Kamei, Kenju

Durham University

25 September 2016

Information Disclosure and Cooperation in a Finitely-repeated Dilemma: Experimental Evidence

Kenju Kamei*

Department of Economics and Finance, University of Durham,
Mill Hill Lane, DH1 3LB, United Kingdom.

This version: September, 2016

Abstract: A large volume of theoretical and experimental studies have suggested that making information on people's past behaviors visible to others may lead to the evolution of cooperation in finitely-repeated environments. But, do people endogenously cooperate with randomly-matched peers by revealing their past when they have an option to hide it? This paper experimentally shows that cooperation does not evolve in a random-matching environment because a large fraction of people do not choose to reveal their past behavior. However, when a costly sorting mechanism (where disclosers are matched with other disclosers; and likewise non-disclosers with other non-disclosers) is present, a stable number of subjects decide to costly disclose their past to join the reputation community and cooperate with other disclosers. Our study at the same time shows that when the sorting process is free, the high efficiency in the reputation community decreases as strategic subjects tend to join the reputation community and attempt to exploit cooperators. These findings suggest an important role of costly sorting mechanisms in the formation of communities (including online platforms) in order for people to sustain a high level of cooperation norms.

Keywords: experiment, cooperation, finitely-repeated dilemma, repeated games, reputation

JEL code: C92, D74, D83

* Email: kenju.kamei@gmail.com, kenju.kamei@durham.ac.uk.

1. Introduction

How people can successfully cooperate in dilemma situations is one of the oldest and the most important questions in economics. One of the actively studied dilemmas in the last few decades by scholars, including social scientists and biologists, is two-person dilemmas. In a two-person dilemma game, the total payoff amount of two players is maximized when they both choose cooperation. However, under the assumption that players are self-interested and that they believe all of their peers are also self-interested, the Pareto-efficient outcome cannot be achieved if it is only finitely repeated. This is because defecting on one's partner results in a higher payoff for the defector, no matter what actions the partner takes.

Scholars have long been interested in how to sustain cooperation in such finitely-repeated two-person dilemmas with random matching. As discussed in the seminal work by Kreps *et al.* (1982), mutual cooperation is theoretically possible if people believe that their peers are not selfish or they merely believe that some of their peers believe that not everyone is selfish. Nevertheless, it is empirically known that although some people attempt to cooperate in earlier periods of finitely-repeated dilemma games, people's level of cooperation steadily declines over time if no additional institutions are available (e.g., Dal Bó and Dal Bó 2014, Kamei 2016a).[1] However, a large volume of studies have proposed that reputation, or information on people's past behaviors, may help people resolve dilemmas and achieve mutual cooperation in various setups (e.g., Engelmann and Fischbacher 2009, Bolton *et al.* 2004 and 2005, Kamei and Putterman forthcoming, Nowak *et al.* 2000, Milinski *et al.* 2002).[2] Research so far has indicated that three particular factors come into play in reputation mechanism. First, the presence of a reputation mechanism may give people material incentives to build a cooperative reputation and to cooperate (e.g., Andreoni and Miller 1993, Kamei and Putterman forthcoming, Bolton *et al.* 2004, 2005, Engelmann and Fischbacher 2009). Maintaining a mutually cooperative relationship

[1] Also see Ledyard (1995) and Chaudhuri (2011).
[2] The impact of exogenously given information on boosting cooperation has also been demonstrated in experiments with infinitely-repeated dilemma games (e.g., Camera and Casari 2009, Kamei 2015).

is more materially beneficial than exploiting counterparts and then being trapped in a mutual defection if they meet more than once. Thus, a selfish individual may mimic the behavior of cooperative individuals and strategically build a reputation as a highly cooperative person in order to establish mutually cooperative relationships with others aided by reputation institutions. Second, some individuals may receive psychological satisfaction from mutual cooperation, explained by concepts such as altruism, inequity aversion and direct reciprocity, while they are averse to being exploited by others (see Fehr and Schmidt (2006) and Sobel (2005) for a survey). As information on the past may serve as a signal that they will cooperate or defect with their peers in the future, reputation mechanisms can act as coordination devices for people to fulfill their desires to cooperate. Lastly, some individuals may even exhibit indirect reciprocity towards strangers, as shown by Engelmann and Fischbacher (2009), Seinen and Schram (2006), Nowak and Sigmund (1998a, b), Wedekind and Milinski (2000), and Wedekind and Braithwaite (2002). Because pro-social behavior can be rewarded by strangers, individuals – even selfish ones, may have a sufficient incentive to cooperate when their behavior can be seen by others in the future. Nevertheless, do people in fact choose to disclose their past in a random-matching community and voluntarily cooperate with each other? Moreover, if subjects' disclosure decisions determine with whom they interact with, i.e., if there is sorting through disclosure in that disclosers are matched with disclosers and non-disclosers are matched with non-disclosers, how do subjects' disclosure decisions change in two-person dilemmas with random matching? Although there is a rich body of studies on the role of reputation institutions ('exogenously' given to subjects) on the evolution of cooperation, little attention has been paid to the possibility where people create reputation communities through voluntary disclosure of information in finitely-repeated dilemma situations, or to people's preferences between an environment with and without a reputation mechanism in dilemma situations.[3]

[3] One exception is the study by Gong and Yang (2014), where subjects can request information about their matched partners' past actions in a finitely-repeated prisoner's dilemma game with random-matching. The request can be made for free. Gong and Yang (2014) find that such an information request device helps subjects enhance mutual

There are many real-world situations in which individuals make information disclosure decisions in random-matching environments. Examples include online-based interactions such as social networks. While some users use their real name and disclose their background, other users only use anonymous usernames on such network. Another example is companies' disclosure of accounting and/or corporate information in annual reports: some voluntarily disclose more detailed information than required by the government. When a firm is looking for a business partner, it may encounter some firms that voluntarily disclose more precise corporate information than others. Situations where individuals make sorting decisions mentioned above have also substantially increased in our modern societies. Examples include online-based interactions, such as dating services and emerging businesses based on the sharing economy. For instance, singles may seek to join an online dating service where the users can see other users' detailed background information, such as incomes (with supporting documents to verify the information in some platforms). These services, however, often require users to pay membership fees or require them to spend some time registering their detailed profiles that will be made available to others. Thus, joining in a network with personal data related to people's reputations is often not cost-free. By contrast, there are usually alternative networks that do not require users to pay or spend time in submitting their detailed background information. This kind of network creates more anonymous interactions across users. Similarly, when deciding on a residential location, one may consider choosing to relocate to an area with well-functioning public monitoring. But in order for a person to become a member of a community with such reputation mechanisms, the person must become known by other members of the community. For this purpose the person will need to incur some cost (e.g., time) to connect with others and introduce him or herself (background, personality, etc.).

---

cooperation, compared with when no information is available. Our study is different from theirs in that subjects need to disclose their past in our study so that the matched partners can observe the past action choices.

According to the past studies on repeated supergames and on the impact of signaling, cooperation can be sustained at a high level through information disclosure or through sorting in a finitely-repeated environment with random matching. For instance, Selten and Stoecker (1986), Andreoni and Miller (1993), and more recently Kamei and Putterman (forthcoming) provide experimental evidence that when subjects repeat finitely-repeated dilemma games (supergames) they may learn to cooperate from supergame to supergame although earlier unraveling of cooperation is also observed in later supergames. In our study, as will be explained in Section 2, we do not let subjects repeat finitely-repeated dilemma games since we are interested in the impact of information disclosure or sorting without learning effects across supergames. Nevertheless, because information disclosure or sorting opportunities may serve as coordination devices, people may be able to sustain cooperation at a high level as an outcome of rational behavior. The literature on voting or cheap talk has demonstrated that opportunities to vote in collective decision-making (e.g., Tyran and Feld 2006) and to send messages (e.g., Duffy and Feltovich 2002, 2006) may help people cooperate with each other in dilemma situations.

We conducted a finitely-repeated two-person public goods game experiment where subjects have an option to disclose their last-period contribution amounts to their matched partners in every period. There are two dimensions in the experimental design. The first dimension is the matching protocol: whether subjects are randomly matched with each other in every period regardless of their disclosure decisions, or subjects are in principle assured that a discloser will be matched with another randomly-selected discloser and likewise a non-discloser will be matched with another non-discloser in each period. The second dimension is the disclosure cost: whether disclosure/sorting is costly or cost-free. Thus, our experimental design is a 2×2 design.

Our experiment shows that cooperation does not evolve only through information disclosure in the random-matching environment. Regardless of the disclosure costs, the average contributions decline gradually and steadily from period to period. A close look at the data

reveals that the unstable cooperation is caused by three features. First, a non-negligible fraction of the subjects choose not to disclose their past and behave opportunistically. Second, subjects correctly anticipate the low disclosure rate in the communities and expect low contributions from non-disclosers. The disclosure rate in the communities is such a low level that building reputations through strong contribution behavior is not materially beneficial for subjects. Third, subjects employ discriminating strategies in that they contribute significantly less when matched with non-disclosers than when matched with disclosers. Such discriminating strategies allow the spread of free-riding behavior within the communities.

However, our data also shows that if subjects are given an opportunity to choose environments between one with the reputation mechanism and another without it, they are able to successfully cooperate with each other in the community with the reputation mechanism. Furthermore, the level of cooperation is higher when sorting into the community with reputation is costly than when it is cost-free. The lower efficiency with free sorting opportunities is caused by the fact that some subjects with malicious intentions to exploit high contributors are more likely to join the community with the reputation mechanism when joining the community is cost-free. This suggests an important role of costly sorting mechanisms for people to achieve a high level of cooperation with their peers.

The main contributions of our paper to the literature are two-fold. First, this paper contributes to a growing body of experimental literature on the evolution of cooperation in finitely-repeated dilemma games. Seminal work by Kreps *et al.* (1982) proposes a possibility of 'rational' cooperation in a finitely-repeated dilemma setting. Experimental work has shown that such rational cooperation may emerge if people *repeat* finitely-repeated dilemma games (Selten and Stoecker 1986, Andreoni and Miller 1993, and Kamei and Putterman forthcoming). However, it is still unclear whether people attempt to build good reputations in a finitely-repeated dilemma without repeating supergames. The present paper explores whether such rational cooperation emerges through people's voluntary disclosure of their past behavior alone.

The results show that if such a disclosure device is the only institution that is available, cooperation may not evolve in random-matching communities. However, we show that when people can self-select an environment, either one with or without the reputation mechanism, cooperation can be sustained at a high level in the community with reputation. This paper not only provides the evidence but also quantitatively explains the driving forces behind the evolution of cooperation with sorting using subjects' beliefs.

The rest of the paper proceeds as follows: Section 2 summarizes the experimental design, Section 3 discusses related literature and hypotheses of our study, Section 4 reports results and Section 5 concludes.

2. Experimental Design

Our study is based on an incentivized laboratory experiment. The design frame used is a two-person public good game (e.g., Ledyard 1995, Chaudhuri 2011). The experiments consist of 20 periods of interactions. In each period, each subject is paired with another subject, is given an endowment of 20 ECUs (experimental currency units) and simultaneously decides how many ECUs they wish to contribute for their pair's joint account.

Each period, except period 1, consists of two stages (Figure 1). In the first stage of a given period $t \in \{2, 3, \ldots, 20\}$, subjects decides whether to disclose their period $t - 1$ allocation decisions to their current partners matched in period t. We vary the treatments by two dimensions (Table 1). The first dimension is the size of disclosure cost: either the disclosure is free or one ECU is charged. In costly-disclosing treatments, one ECU is deducted at the end of a given period (a subject has 20 ECUs in her allocation-decision stage even when she decides to costly disclose her last-period allocation amount). The second dimension is the matching protocol: one that a discloser is assured that he or she is matched with another discloser, or the other that each subject is randomly matched with another subject, regardless of disclosure decisions (Section 2.1). Our design is therefore a 2×2 factorial design. The four treatments are named as the

"Costly Sorting" treatment, abbreviated as the C-Soring treatment, "Free Sorting" treatment, abbreviated as the F-Soring treatment, the "Costly Disclosure, Random Matching" treatment, abbreviated as the C-RM treatment, and the "Free Disclosure, Random Matching" treatment, abbreviated as the F-RM treatment.

2.1. Matching Protocol

In period 1, each subject is randomly matched with another subject without making any disclosure or sorting decisions; and then plays the public goods game with the matched subject.

As mentioned, each period after period 1 consists of two stages. In the two sorting treatments (F-Sorting, C-Sorting), subjects in period $t \in \{2, 3, \ldots, 20\}$ decides whether to disclose their period $t - 1$ allocation amounts to join the 'reputation' community in period t. If a subject chooses to disclose her behavior in the previous period, she is randomly matched with another subject that likewise chose disclose his last-period contribution amount. The disclosing cost is one ECU in the C-Sorting and is free in the F-Sorting treatment. Alternatively, a subject could join a community of non-disclosers in period t, which we call the 'anonymous' community, by choosing not to disclose her allocation amount in the previous period. In that case, her previous contribution amount would not be informed to her current matched partner (and one ECU is not deducted at the end of period t in the C-Sorting treatment). She is then randomly matched with another non-discloser; and the partner is only informed that he is randomly matched with another person in the community of non-disclosers.[4]

In the two random-matching treatments (C-RM and F-RM), each subject in period $t \in \{2, 3, \ldots, 20\}$ decides whether to disclose his or her period $t - 1$ allocation amount in the period t interaction as in the C-Sorting and F-Sorting treatment, respectively. However, in these two

[4] If the number of disclosers (non-disclosers) is an odd number, one discloser is randomly matched with a non-discloser. This event happened only around 9.3% of the paring in the C-Sorting treatment and 6.9% of it in the F-Sorting treatment. Further data analyses (Section 4) indicate that the paper's findings are robust, regardless of whether we use data of pairs consisting of pairs with the same preferences (two-disclosers pairs and two-non-disclosers pairs) only or all data.

treatments, each subject is randomly matched with another subject, regardless of their disclosure decisions. The consequences of subjects' disclosure decisions, except the matching procedures, are the same as those in the Sorting treatments. That is, when a subject decides to disclose in period t, the subject's matched partner in that period is informed of her period $t-1$ allocation decision; and the partner is not given this information when the subjects decides not to disclose.

There are no subject identification numbers provided during the entire experiment. Thus, the only way that subjects could form reputation is through their disclosure (sorting) decisions in the C-RM and F-RM (C-Sorting and F-Sorting) treatments.

2.2. Allocation Decisions

Each matched pair in every period plays a two-person linear public goods game. Only integers between 0 and 20 are allowed for their contribution amount. The payoff consequences are as follows: a subject receives one ECU for each ECU she allocates to her private account. By contrast, if she contributes one ECU to her joint account, she and her partner each receive 0.8 ECUs, which is less than one ECU, from the joint account. However, the total group payoff is maximized when two subjects in a pair both contribute all of their endowments to the joint account ($0.8 \times 2 = 1.6 > 1.0$). Suppose that subject i contributes $c_{i,t}$ to her joint account in period t. Then, subject i obtains the following payoff in period t:

$$\pi_{i,t} = 20 - c_{i,t} + .8 \cdot (c_{i,t} + c_{j,t}). \quad (1)$$

Here, subject j is subject i's matched person in period t and $c_{j,t}$ is subject j's contribution to the joint account in period t.

2.3. Conditional Contribution Schedule and Beliefs

We include two kinds of additional tasks in order to explore driving forces behind subjects' disclosure decisions in the C-RM and F-RM treatment and subjects' community formation in the C-Sorting and F-Sorting treatments. The first additional task is elicitation of beliefs. First, in each allocation-decision stage (period 1 and second stage of period $t \in \{2, 3, \ldots,$

20}) of all treatments, subjects are asked about their beliefs on their matched partner's contribution amount in a given period. Elicited beliefs are used in order to examine the role of subjects' beliefs on the evolution of cooperation or on community formation (e.g., Kreps *et al.* 1982). This elicitation task is not incentivized in order to avoid a hedging problem.[5] Second, subjects in the C-RM and F-RM treatments are also asked to answer the expected number of disclosers (except themselves) in period *t*. This elicitation task is also not incentivized.

Second, we elicit subjects' cooperation types using Fischbacher *et al.* (2001). Specifically, each subject is asked to answer how many ECUs they wish to allocate to their group, conditional on each of the other group members' average contributions. This task is incentivized. The details of the procedure for this part are provided in the online Appendix A. This task is conducted before the 20 periods of the finitely-repeated public goods game. However, subjects are informed of the outcomes of this task only after they complete the 20 periods of the public goods games in order to minimize the effects of this task on their behavior. In addition, group composition is randomly changed between this elicitation task and the 20 periods of repeated dilemmas. The elicited conditional contribution preferences are used to examine the possibility that cooperation-oriented types are more likely to disclose their past behavior.

2.4. Experimental Procedure

All experiments except the instructions were programmed using z-tree (Fischbacher 2007). All eligible subjects were sent solicitation messages via ORSEE (Online Recruitment System for Economic Experiments) developed by Greiner (2015); and subjects voluntarily registered for and participated in the experiment. No subjects participated in more than one session. At the onset of the experiment, neutrally framed instructions for the conditional contribution task were handed out to subjects and read aloud by the experimenter. Once the task for eliciting cooperation types was over, instructions for the main part of the experiment (20

[5] In the experiment, this task was only included in the instructions shown on the computer screens, not in the hard copy of instructions distributed to subjects, in order to avoid making this task salient.

rounds of public goods games), which are likewise neutrally framed, were distributed and read aloud. Control questions were included in each set of the instructions so as to check the subjects' understanding of the experiment. Communication was prohibited during the entire experiment. Subjects were also asked to switch off any electronic devices (e.g., cell phones) during the experiment. Subjects were privately paid based on their accumulated ECUs (40 ECUs in the experiment are exchanged for £1 of real money) at the end of the experiment.

3. Related Literature and Hypotheses

Standard theory predicts a point estimate in our environment because the MPCR is 0.8 (see Eq. (1)). That is, contributing zero ECUs to the joint account is a strictly dominant strategy for each subject in any given period ($\partial \pi_{i,t}/\partial c_{i,t} = -.2 < 0$). Therefore, with the logic of backward induction, each subject would contribute nothing to their joint account in every period, assuming that they believe that other subjects always choose defection without considering their opponents' reputations. Considering the peers' uniform full free-riding behavior, no one would costly disclose their past towards the matched person in any period in the C-RM treatment; and likewise no one would costly sorts into the reputation community in the C-Sorting treatment, in order to save a disclosure/sorting cost. Disclosure or sorting decisions do not affect subjects' payoffs in the F-RM and F-Sorting treatments, respectively, because these decisions can be made for free and their peers would select defection always; hence, subjects would randomly decide whether to disclose and sort in the two treatments.

HYPOTHESIS 1: Standard Theory Prediction.

(a) No one costly discloses their last-period contribution behavior in the C-RM treatment. No one costly sorts into the reputation community in the C-Sorting treatment. (b) Disclosure and sorting decisions are randomly made by subjects in the F-RM and R-Sorting treatments, respectively. (c) Subjects contribute nothing to the joint accounts in each period in all treatments.

A large body of experimental research has partially confirmed Hypothesis 1(c) in finitely-repeated dilemma games when there are no institutions such as disclosure and sorting involved. It demonstrates that although subjects contribute around 40% to 60% of the endowment in the public goods games and around 30% to 40% of subjects choose to cooperate in the prisoner's dilemma games in earlier periods, they decrease the levels of cooperation steadily from period to period (e.g., Ledyard 1995 and Chaudhuri 2011 for surveys on public goods games; Andreoni and Miller 1993, Dal Bó and Dal Bó 2014, Dal Bó *et al.* 2010 for evidence on prisoner's dilemma games).

But on the other hand, Selten and Stoecker (1986), Andreoni and Miller (1993), and also recently Kamei and Putterman (forthcoming) have shown that subjects may be able to achieve high levels of cooperation in earlier periods of finitely-repeated dilemma setups if they repeat a finitely-repeated dilemma game (supergame) under some conditions.[6] In Selten and Stoecker (1986), subjects repeated a 10-periods finitely-repeated prisoner's dilemma game 25 times where the pairing stayed fixed within supergames, but randomly changed across the supergames. The subjects tended to be more cooperative in the first period of a given supergame as they gained experiences, but end-game defection at the same time shifted to earlier periods in later supergames. Andreoni and Miller (1993) likewise let subjects play 10-periods finitely-repeated prisoner's dilemma games 20 times. In the treatments where a partner-matching protocol was used within supergames but a different partner was assigned to a subject in each supergame, the subjects gradually raised the levels of cooperation in earlier periods from supergame to supergame while the timing of end-game defection shifted to earlier periods. In later supergames, the subjects' within-supergame cooperation dynamics became stable.[7] Andreoni and Miller (1993) also conducted treatments where a random-matching protocol was used within

[6] See Hauk and Nagel (2001) also.

[7] Cox *et al.* (2015) showed that subjects learn to strategically cooperate from supergame to supergame when the second movers are informed of the first movers' history from earlier supergames and likewise when the first movers are informed the second mover's history from earlier supergames. In Cox *et al.* (2015), a partner-matching protocol was used within a supergame and a perfect random-matching protocol was used across the supergames as in Andreoni and Miller (1993).

supergames. Their results from the random-matching sessions indicated that (a) the average cooperation rate was much lower than when the partner matching was used and (b) unlike the partner-matching treatment, the average cooperation rate in the first five periods was not different from that in the last five periods.[8] Nevertheless, even in the later supergames with random matching, the average cooperation rate was not zero as a small number of subjects still cooperated in earlier periods. Kamei and Putterman (forthcoming) furthermore showed that when subjects can choose with whom they interact and also when they can observe information on peers' past action choices within a given supergame, cooperation can grow from supergame to supergame. The subjects' learning behavior in later supergames in Selten and Stoecker (1986), and the increase in cooperation with the partner-matching treatment in Andreoni and Miller (1993) and Kamei and Putterman (forthcoming) can be rationalized by assuming that some subjects have non-selfish preferences (e.g., Fehr and Schmidt 2006, Sobel 2005) or some subjects believe that some peers are not selfish or believe that some peers believe that not everyone is selfish (Kreps *et al.* 1982). Kreps *et al.* (1982) theoretically demonstrate a possibility of such 'rational' cooperation behavior in a finitely-repeated dilemma game assuming that some people believe that some fraction of their peers act on the tit-for-tat strategy.[9] Andreoni and Miller (1993) and Kamei and Putterman (forthcoming) also provide empirical evidence that some subjects are non-selfish types and its presence may contribute to the above findings.[10]

Our experiment uses a standard single-supergame design. In addition, a random-matching protocol is used within communities. Recall a low level of cooperation with random matching in Andreoni and Miller (1993). Nevertheless, subjects' rational cooperation behavior seen in the past studies may happen even with our standard design because of the disclosure (sorting)

[8] The mean periods in which a subject selected defection were very small: they were constantly between 2 and 3 over the supergames in the random-matching sessions (Fig.4 in Andreoni and Miller 1993).
[9] We note that strategies like the 'tit-for-tat' can be evolutionarily stable and can lead to mutual cooperation also in infinitely-repeated interactions (Axelrod and Hamilton 1981, Axelrod 1984, Wedekind and Milinski 1996).
[10] See Reuben and Suetens (2012) also. They used an indefinitely-repeated sequential prisoner's dilemma game with strategy method and showed that a non-negligible fraction of second movers select to cooperate if the game does not end in a given period and matched first-movers selected cooperation. The prevalence of non-selfish types has also been shown in past studies (e.g., Fischbacher *et al.* 2001, Kurzban and Houser 2005).

institutions in the F-RM and C-RM (F-Sorting and C-Sorting) treatments as they may act as coordination devices among subjects. The effectiveness of signaling has been shown in both public goods game and prisoner's dilemma game setups (e.g., Tyran and Feld 2006, Duffy and Feltovich 2002).[11] For an example of public goods games, Tyran and Feld (2006) show that collectively selecting a non-deterrent sanction rule in a one-shot public goods dilemma may encourage subjects to cooperate more, compared with when the same rule is exogenously imposed, because voting for sanctions can serve as a signal that subjects would cooperate with others. For an example of prisoner's dilemma games, Duffy and Feltovich (2002) show that a cheap-talk opportunity (where one subject in a pair can send a costless, nonbinding message to the other subject in the pair as to which action he or she will select) in a prisoner's dilemma game makes cooperation more likely.[12] Given that disclosure and sorting opportunities can serve as devices to send signals, some subjects may choose to disclose or sort into the reputation community and then contribute larger amounts than non-disclosers, as in past studies with voting, cheap talk and communication.

HYPOTHESIS 2: *Some subjects disclose their past towards their matched persons in the C-RM and F-RM treatments. Some subjects sort into the reputation community through disclosure in the C-Sorting and F-Sorting treatments.*

Precisely, how many subjects disclose their last-period action choices in the C-RM and F-RM treatments as in Hypothesis 2? What about for the C-Sorting and F-Sorting treatments? How likely are contributions to be sustained at high levels with the disclosure institution or with the sorting institution? A well-known strategy, similar to the tit-for-tat, in the context of public goods games is the so-called 'conditional cooperation strategy' (e.g., Fischbacher *et al.* 2001,

[11] Giving an opportunity to subjects so that they can send signals also improve coordination among subjects in coordination games (e.g., Cooper *et al.* 1992 and Kamei 2016b).

[12] Also see Duffy and Feltovich (2006), which show that the cooperation rate does not rise furthermore when receivers of message can also see the senders' previous-round actions, compared with when only the cheap-talk opportunity is available.

Fischbacher and Gächter 2010).[13] For an illustration purpose, we will assume that subject i is a material payoff maximizer. First consider subject i's decisions in period t, where $t < 19$, in the C-RM and F-RM treatments. Suppose that subject i contributed a large amount c to the joint account in period $t - 1$. Suppose also that subject i believes that p percent of her peers are disclosing their past and are acting on the conditional cooperation strategy where they contribute amounts that are the same as their pair partners' last-period contributions in a given period if the partners' last-period behaviors are observable; and they contribute zero ECUs, otherwise. Suppose also that the rest ($1 - p$ percent of their peers) are acting on the "always defect" strategy according to her belief. With the "always defect" strategy, players unconditionally select defection. Lastly, suppose that subject i correctly anticipates their peers' behavior and all of her above beliefs turn out to be true. We first study the case where subject i disclosed her past also in period t. Under this setup, if p is sufficiently large, it is materially beneficial for i to continue to contribute a large amount in period t. To see this, suppose that subject i chooses to contribute the large amount c again in period t. Under this situation, i obtains a payoff of: $20 - c + r \cdot c \cdot [1 + (p/100)]$, where $r = .8$. Subject i's payoff from the two consecutive periods (periods t and $t + 1$), $V_{i,t}$, is calculated as:

$$V_{i,t} = -1_{\text{C-RM}} + 20 - c + r \cdot c \cdot [1+(p/100)] + (1/2) \cdot V_{t+1},$$

where $1_{\text{C-RM}}$ is an indicator variable which equals 1 for the C-RM treatment; and 0, otherwise. From the condition: $V_{i,t} = V_{t+1}$, we can find the value of $V_{i,t}$:

$$V_{i,t} = -2 \cdot 1_{\text{C-RM}} + 40 + \{-2 + 2r \cdot [1+(p/100)]\}c.$$

By contrast, if subject i chooses to contribute 0 in period t, subject i's maximum payoff in periods t and $t + 1$ would become:

$$V'_{i,t} = -1_{\text{C-RM}} + 20 + r \cdot c + 20 = -1_{\text{C-RM}} + 40 + r \cdot c.$$

Therefore, we have, by using $r = .8$:

[13] Some subjects' conditional cooperation behavior can be rationalized by other-regarding preferences, such as inequity aversion (e.g., Fehr and Schmidt 1999) and reciprocity (e.g., Rabin 1993, Dufwenberg and Kirchsteiger 2004, Cox *et al.* 2007).

$$V_{i,t} - V'_{i,t} = -1_{\text{C-RM}} + \{-1.2 + 1.6\,(p/100)\}c. \quad (2)$$

This suggests that $V_{i,t}$ is bigger than $V'_{i,t}$ if p is large enough that $p > \underline{p} = 75[\%]$ in the F-RM treatment; and that the former is bigger than the latter if $p > 75 + 62.5/c[\%]$ in the C-RM treatment. Note that $75 + 62.5/c$ is less than 100 if $c > 3$ (whose condition is very weak and would be most likely met assuming the findings of past experimental research).

We now compare subject i's payoff between when i disclose her past and when i does not do so. An analysis indicates that it is not materially beneficial for i to not disclose if p is sufficiently large because such hiding triggers her current-period partner's defection. The threshold value of p that makes disclosing materially beneficial can be calculated by comparing $V_{i,t}/2$ and 20 (which is the defection payoff):

$$V_{i,t}/2 - 20 = -1_{\text{C-RM}} + \{-0.2 + 0.8\cdot(p/100)\}c > 0. \quad (3)$$

Condition (3) holds for $p > 25[\%]$ in the F-RM treatment and for $p > 25 + 125/c[\%]$ in the C-RM treatment. Here, $25 + 125/c$ is less than 100 unless $c = 0$ or 1 ($c = 0$ or 1 is very unlikely situations based on the evidence from past experimental research).

All of the aforementioned calculations indicate that when p is sufficiently large, cooperation can be sustained at a high level only with selfish motives. These also indicate that such a high level of cooperation is more likely to be achieved in the F-RM treatment than in the C-RM treatment because the threshold p is lower in the former treatment than in the latter. We note that in these considerations, subjects do not need to exhibit non-standard preferences as discussed in Kreps *et al.* (1982). If a sufficiently large fraction of subjects believe that a large percentage of their peers will disclose their past and choose (mimic) such conditional cooperation strategy, then rational cooperation can happen even though all are non-selfish types.

In the C-Sorting and F-Sorting treatments, each discloser is assured that he or she is matched with another discloser (except some small cases mentioned in Section 2). This means that p is close to 100% in the reputation community, assuming that joining the reputation community is a signal of subjects' conditional cooperation behavior as in Hypothesis 2. Thus, the

above discussion on the threshold values of p which make cooperation materially beneficial suggest that rational cooperation can be more easily achieved in the reputation community of the C-Sorting than in the C-RM treatment, and likewise in the F-Sorting treatment than in the F-RM treatment. All of these considerations are summarized as Hypothesis 3 below.

HYPOTHESIS 3: *(a) If subjects believe that more than 75[%] (75 + 62.5/c[%]) of their peers are disclosers and are employing the conditional cooperation strategy, then cooperation can be sustained at a high level in the F-RM (C-RM) treatment. (b) Cooperation is sustained at a higher level in the F-RM treatment than in the C-RM treatment. (c) Cooperation is more likely to be sustained at a higher level in the reputation community in the F-Sorting (C-Sorting) treatment than in the F-RM (C-RM) treatment. (d) Cooperation is sustained at a higher level in the reputation community than in the anonymous community in the two sorting treatments.*

We note that mutual cooperation may be more likely to be achieved in the C-Sorting and F-Sorting treatments than in the C-RM and F-RM treatments, respectively, even if there are no material incentives to build cooperative reputations, if (i) some subjects are non-selfish types and (ii) such non-selfish types are disclosers. As briefly mentioned earlier, there is a large volume of literature on the prevalence of non-selfish human types. It has shown that people may indeed be able to sustain cooperation if they are matched with like-minded others with respect to the degree of cooperativeness (e.g., Gunnthorsdottir *et al.* 2007, Gächter and Thöni 2010). Therefore, testing Hypothesis 3 is not a simple task. In order to disentangle subjects' selfish reputation-building motives from non-selfish motives, we classify subjects' cooperation types by using the conditional contribution task (Fischbacher *et al.* 2001) included at the onset of the experiment; and we then study the reputation building behavior for each cooperation type (see Section 4).

Hypothesis 3(a) can be explored by using the elicited beliefs on the percentage of disclosers in the C-RM and F-RM treatments. If cooperation evolves and is sustained at high levels in the C-RM and F-RM treatments with the logic discussed above, then the average beliefs

should be at least greater than $75 + 62.5/\bar{c}$ [%] and 75[%] and in the C-RM and F-RM treatments, respectively, where $\bar{c}$ is the average contribution of disclosers in the C-RM treatment.

Hypotheses 3(c) and (d) imply that the fractions of subjects who disclose to sort into the reputation community in the F-Sorting and C-Sorting treatments are higher than the fractions of subjects who disclose in the F-RM and C-RM treatments, respectively, because achieving mutual cooperation would be easier in the reputation community than in the community where disclosers and non-disclosers co-exist. This leads to the following additional hypothesis.

HYPOTHESIS 3(e): *The percentage of subjects that sort into the reputation community through disclosure in the C-Sorting (F-Sorting) treatment is higher than that of subjects that disclose in the C-RM (F-RM) treatment.*

Lastly, we note that our paper is also related to a large body of experimental literature on (i) endogenous group formation (e.g., Ahn *et al.* 2008 and 2009, Aimone *et al.* 2013, Bayer 2011, Coricelli *et al.* 2004, Ehrhart and Keser 1999, Gallo and Yan 2015, Kamei and Putterman forthcoming, Page *et al.* 2005) and (ii) institutional choices based on voting with feet (e.g., Fehr and Williams 2013, Grimm and Mengel 2009, Gürerk *et al.* 2006, Rockenbach and Milinski 2006) for the C-Sorting and F-Sorting treatments. As for literature (i), it proposes that cooperation may evolve in public goods games or prisoner's dilemma games with fixed group size if individuals are provided with both an ability to choose with whom they interact and sufficient information on other players' past behaviors (e.g., Page *et al.* 2005, Bayer 2011, Kamei and Putterman forthcoming).[14] For instance, subjects in Page *et al.* (2005) formed groups of four in every three periods using a mutual selection protocol and played four-person public goods games with each other. Page *et al.* (2005) showed that the endogenous group formation is helpful in sustaining cooperation, compared with the control treatment without such endogenous

[14] Also see Aimone *et al.* (2013).

re-grouping (also see Kamei and Putterman, forthcoming).[15] Bayer (2011) let subjects play six sets of four periods in sequence, where each subject had a endogenous matching protocol based on Gale and Shapley (1962) at the onset of each set and then played two-person public goods games with each other for four periods in a given set. He found that the endogenous group formation increases efficiency (contribution and payoffs). Similar findings have also been seen in dilemma games with variable group size if entry to a new group requires the group members' agreement (e.g., some treatments in Ahn *et al.* 2008 and 2009, Charness and Yang 2014, Gallo and Yan 2015). For instance, Ahn *et al.* (2008) let subjects play a sequence of endogenous group formation followed by a public goods game with negative externality in each group, for 25 times. Ahn *et al.* (2008) found that for a given group size, subjects contribute more and they suffer less from the congestion problem when group members' approval is required to join a group than otherwise. In Charness and Yang (2014), each group was periodically given an opportunity to merge with another group on condition that at least 60% of members in each group wished to do so. Their data showed that such group formation opportunities increase subjects' contribution amounts, compared with when subjects are exogenously assigned to a group with the same fixed group size. In a dynamic network formation experiment where a new link can be formed when the recipient accepts a proposal made by another subject, Gallo and Yan (2015) found that global reputation information (a list of past actions of every subject) can improve the efficacy of dynamic network formation. On the other hand, however, when allowed to change their interaction groups at will (i.e., without their members' agreement for moving-in or out, or without threat of exclusion), it is known that selfish individuals may attempt to join groups that have cooperators, aiming to exploit them, and as a result cooperation may be difficult to evolve in N-person dilemmas, where group size $N > 2$ (e.g., some treatments in Ahn *et al.* 2008 and 2009, Ehrhart and Keser 1999). For example, Ehrhart and Keser (1999) let subjects play a

[15] We note that the effectiveness of endogenous group formation may be affected by its group formation procedure. For instance, the experiment in Coricelli *et al.* (2004) indicated that unidirectional partner selection may be more effective in increasing voluntary contribution than bidirectional partner selection.

sequence of group adherence decision followed by a public goods game. In the group adherence decision stage, each subject was given information including the list of group's identification numbers and average group contributions, and was then given an opportunity to switch groups. The experiments indicated that cooperation was not sustained at a high level because groups with high contributors were joined by selfish individuals with intentions to exploit them. The findings from past experiments such as Ehrhart and Keser (1999) may or may not be extended to our experimental setup. In our experiment, unlike these past studies on endogenous group formation, subjects do not engage in an *N*-person dilemma game with all group members (*N* persons). Instead, subjects who decide to join the reputation or anonymous community are randomly matched with another subject within their community and they then play the *two*-person prisoner's dilemma game with each other. Thus, although such chasing behaviors of malicious individuals may be prevalent, subjects are able to adopt tit-for-tat-like discriminating strategies targeted to the specific malicious members with help of the reputation mechanisms, as discussed with the conditional cooperation strategy above. As such, chasing behaviors by malicious individuals may be less beneficial in our study and cooperation may be sustained at a high level in the reputation community of the sorting treatments. Nevertheless, such a punishment or protective action lowers a member's own reputation. Accordingly, there is also a chance where small instances of low contribution events may spread to other members and may quickly destroy the community's cooperation norms. As for a comparison between the C-Sorting and F-Sorting treatments, the average contribution in the reputation community could be higher in the C-Sorting treatment because the positive cost of sorting may discourage subjects with such malicious intentions from joining the reputation community to some degree.

HYPOTHESIS 4: *Cooperation in the reputation community is sustained at a higher level in the C-Sorting treatment than in the F-Sorting treatment.*

As for literature (ii) mentioned above, for instance, Gürerk *et al.* (2006) has shown that when given a choice whether to join a group with an informal sanctioning institution (where each

group member has an option to costly punish others *after* observing others' action choices) or to join one without it, almost all subjects select the sanctioning institution and successfully cooperate with each other. Rockenbach and Milinski (2006) show that the interaction between costly punishment and reputation building enhances cooperation. Specifically, they let subjects choose a stage game between (a) a public goods game followed by costly punishment stage and an indirect reciprocity game and (b) a public goods game followed only by an indirect reciprocity game, in every period. They found that regime (a) was more popular than regime (b) in later periods. The data of Rockenbach and Milinski (2006) also indicated that those who selected regime (a) contributed significantly more than those who selected regime (b) (the average contributions declined steadily from period to period under regime (b)). Fehr and Williams (2013) let subjects choose among four environments including centralized sanctioning where one member is elected by voting and only the elected member can punish each member in this environment. They found that the centralized sanctioning institution and an informal sanctioning institution are most frequently chosen and subjects can achieve higher cooperation with these institutions.[16] Lastly, Grimm and Mengel (2009) let subjects choose one out of two payoff matrices. With one of the two payoff matrices, a subject would have not only less temptation to defect but also a less mutual defection payoff. The subjects then played an *N*-person prisoner's dilemma game, where *N* is the community size that selected the same payoff matrix. Grimm and Mengel (2009) found that subjects cooperate significantly more when they sacrifice the defector gains. These past studies are different from our study with two important aspects. First, in these past studies, once a subject self-selects an environment in a given period, her payoff depends on action choices of *all* members in the selected environment. In our study, each subject is randomly matched with *one* person in the selected community and then plays the two-person dilemma games. Second, unlike the past studies, subjects in our study are provided an opportunity to join an environment with the reputation mechanism only. That is, except the

[16] See also Nicklisch *et al.* (2015).

presence of the reputation mechanism, not only subjects' set of actions but also their payoff structure is the same between the reputation and anonymous communities (subjects neither have an additional punishment stage nor have a different payoff formula in the reputation community). Thus, for instance, unlike the research such as Gürerk *et al.* (2006), a constituent member must avoid contributing high amounts in order to avoid being exploited, when he encounters with a low cooperator in a given period. This kind of defection behavior may harm the community's cooperation norms as already discussed. The competitive advantage of the reputation environment over the anonymous environment is therefore not clear in our study in the context of these two branches of literature (i) and (ii) unlike Hypothesis 3(d).

4. Results

We conducted 16 sessions – four sessions for each treatment – at Durham University in August 2015, February, June through August in 2016. A total of 180 Durham University subjects participated in the experiment. The session size was 12 subjects in all sessions, except one session each in the C-Sorting, F-RM and F-Sorting treatments.[17] The average payment (including participation fee of £3) was £15.81 with a standard deviation of £0.89. The average duration of the experiment (including payment to the subjects) was around 90 minutes.

4.1. Subjects' Disclosure and Sorting Decisions, and their Average Contributions

We first overview the trends of subjects' disclosure decisions and average contribution amounts in the C-RM and F-RM treatments. Figure 2 reports the trends. The data shows that Hypothesis 1 does not hold. Even when there are no sorting opportunities, a non-negligible fraction of subjects choose to disclose their past as in Hypothesis 2, irrespective of whether the disclosure cost is zero or not. The overall fraction of disclosers in the C-RM treatment is 23.0% (Table 1). The disclosure rate increases significantly in the F-RM treatments, relative to the C-RM treatment (Part (II) of Appendix Table C.2), and the fraction of disclosers is a little above

[17] The session size was eight for these three sessions.

the half of the subjects, 53.9% (Table 1).[18] The subjects' average beliefs on the fraction of disclosers transit almost parallel to the actual percentages of disclosers in the C-RM and F-RM treatments (Panels I(c) and II(c) in Figure 2). Although these beliefs are clearly above zero, they are much lower than the percentages of conditional cooperative disclosers that selfish players must believe so that cooperation is theoretically materially beneficial in the random-matching treatments (Hypothesis 3(a)). The trends of average contributions in the C-RM and F-RM treatments are similar to each other and the average contributions follow the typical free-riding dynamics seen in the literature (e.g. Ledyard 1995, Chaudhuri 2011). Specifically, average contributions begin with around 40% and 50% of the endowment and steadily decline from period to period in the C-RM and F-RM treatments, respectively (Panels I(a) and II(a) in Figure 2). The decreasing trends are statistically significant (columns (1) and (4) in Table 2). This result on the decreasing trends is consistent with the failure of Hypothesis 3(a) in our experiment.

RESULT 1: *Non-negligible fractions of subjects disclose their past in the C-RM and F-RM treatments. Subjects correctly form beliefs on the percentages of disclosers among peers. However, subjects' beliefs on the disclosure rate are much lower than the threshold percentages summarized in Hypothesis 3(a). The average contributions in the C-RM and F-RM treatments steadily decline from period to period.*

When the sorting device is present, significantly larger fractions of subjects disclose their last-period contribution amounts to join the reputation community, compared with the corresponding random-matching treatments.[19] The fractions of those who sorted into the reputation community are 52.8% and 70.7% in the C-Sorting and F-Sorting treatments, respectively (Table 1). The higher disclosure rates in the sorting treatments support Hypothesis 3(e). In addition, Hypothesis 3(d) holds for both the C-Sorting and F-Sorting treatments: the

[18] A regression analysis shows that subjects' frequency of disclosing in each of the C-RM and F-RM treatments is significantly different from 0. See the estimates of the constant terms of Panel (I) in Appendix Table C.2.
[19] See Panel (I) of Appendix Table C.2.

average contributions in the reputation community are higher than those in the anonymous community in the two treatments (Panels I(a) and II(a) in Figure 2).[20]

RESULT 2: *Significantly larger fractions of subjects disclose their past in the C-Sorting and F-Sorting treatments, compared with the C-RM and F-RM treatments, respectively. The average contributions are significantly higher in the reputation community than in the anonymous community in the C-Sorting and F-Sorting treatments.*

A close look at the data indicates that subjects' contribution dynamics in the reputation community differ between the C-Sorting and F-Sorting treatments. On the one hand, disclosers in the reputation community continuously contribute more than half of their endowment, ranging from around 10 to 15 ECUs on average, to their joint accounts in the C-Sorting treatment. The high cooperation norms were well sustained, except the end period (Andreoni 1988). This trend is clearly different from that in the C-RM treatment as well as in the anonymous community in the C-Sorting treatment (Panel I(a), Figure 2). The overall average contribution of the disclosers in the reputation community is 11.96 ECUs, which is significantly higher than the average contribution in the C-RM treatment, 6.02 ECUs (Mann-Whitney test using session-average data, p-value = .021, two-sided).[21] Although the disclosers that joined the reputation community had to pay one ECU for disclosure, their average per-period payoff (26.01 ECUs) was significantly larger than the average per-period payoff in the C-RM treatment (23.40 ECUs) (Mann-Whitney test using session-average data, p-value = .021, two-sided).

On the other hand, in the F-Sorting treatment, where sorting can be made for free, the average contributions in the reputation community hover between 8 ECUs and 11 ECUs during the first ten periods and then between 6 ECUs and 9 ECUs during the second ten periods. As

[20] A session-average Wilcoxon signed ranks test finds that the difference is statistically significant if we use all data of the two treatments (p = .0117, two-sided).

[21] The average contribution in the anonymous community (the set of non-disclosers) in the C-Sorting treatment was almost the same as the average contribution in the C-RM treatment (Table 1, Figure 2I(a)). A Mann-Whitney test finds that the average contribution of the disclosers in the reputation community is also significantly different from that of non-disclosers in the anonymous community in the C-Sorting treatment (p-value = .021, two-sided).

shown in Panel II(a) of Figure 2, the average contributions in the reputation community are on average in a declining trend, unlike the ones in the C-Sorting treatment. A regression analysis confirms that the decreasing trend is significant (Table 2). The overall average contribution in the reputation community was 9.03 ECUs, which was higher than the average contribution in the F-RM treatment, 7.62 ECUs (Table 1), but the difference between the two is not statistically significant (Mann-Whitney test using session-average data, p-value = .2482, two-sided).

As in Result 2, low cooperation norms were commonly observed in the anonymous community in both of the two sorting treatments. Non-disclosers in the C-Sorting treatment contributed always less than 9 ECUs on average and steadily and significantly decreased their contributions from period to period (Panel I(a) in Figure 2, Table 2). In the F-Sorting treatment, non-disclosers contributed low amounts from the earlier periods: the average contributions hovered between 4 ECUs and 8 ECUs throughout the experiment (Panel II(a) in Figure 2).

RESULT 3: *Strong cooperation norms are well sustained in the reputation community in the C-Sorting treatment. Specifically, the average contribution is significantly higher in the reputation community in the C-Sorting treatment compared with the C-RM treatment. This is not true for the reputation community in the F-Sorting treatment, whose cooperation norms are not significantly different from those in the F-RM treatment.*

In short, our data showed that cooperation is not sustained at high levels in the C-RM and F-RM treatments (Result 1), but that the costly sorting mechanism serves as an effective coordination device for subjects to cooperate with each other in the reputation community (Results 2 and 3). But, to what extent did subjects attempt to strategically build a good reputation in the C-RM and F-RM treatments? What drove the high efficiency in the sorting treatments? Why did the efficiency in the reputation community differ between the two sorting treatments as in Result 3? We will closely look at the data to answer these questions in the next two sections.

4.2. Disclosure Decisions and Action Choices in the C-RM and F-RM treatments

We found that subjects were not able to sustain cooperation only through disclosure, regardless of the disclosure costs, in the two random-matching treatments (Result 1). A detailed analysis reveals that subjects' disclosure decisions and their subsequent decisions to contribute are closely linked to each other as assumed in the discussions in Section 3. Figure 3 reports the average contribution amounts by disclosure decision in the C-RM and F-RM treatments. It indicates that disclosers on average contributed 11.0 ECUs and 10.5 ECUs in the C-RM and F-RM treatments, respectively, both of which are significantly higher than the average contribution amounts of non-disclosers in these two treatments (which are 4.4 ECUs and 3.9 ECUs).[22]

RESULT 4: *Disclosers are significantly more likely than non-disclosers to contribute large amounts to the joint accounts in the C-RM and F-RM treatments.*

Result 4, combined with the realized low levels of disclosure rates (Result 1), suggests that it was not materially beneficial for subject *i* to strategically build a cooperative reputation by disclosing his or her past and acting pro-socially in the C-RM and F-RM treatments.

Despite the failure of rational cooperation in the C-RM and F-RM treatments, some selfish subjects may strategically contribute positive amounts to the joint account to build cooperative reputations and then disclose the contribution amounts in the following periods. The conditional contribution schedule (Fischbacher *et al.* 2001) elicited from each subject can be used for this analysis. The average conditional schedule of the overall subject population group falls in line with the standard conditional cooperator type: own contribution amounts are significantly positively increasing in the others' average contribution amounts (Appendix Figure C.2). In our detailed analysis below, we focus on two cooperation types: "conditional cooperators" and "free riders." Similar to Fischbacher *et al.* (2001), we define those whose own contribution amounts and the others' average contribution amounts are significantly positively

[22] We conducted an individual random-effects ordered probit regression in which the dependent variable is subject *i*'s contribution amount ($\in\{0, 1, 2, \ldots, 20\}$) in period *t* and the independent variable is a dummy which equals 1 if *i* disclosed and 0 if *i* did not disclose his or her last-period contribution amount. Standard errors were clustered by session for each treatment. We found that the dummy variable obtains a significantly positive coefficient at the 1% level for each of the C-RM and F-RM treatments. The results are omitted to conserve space.

correlated at least at the 5% level (according to spearman's ρ correlation coefficients) as the conditional cooperators. We also define those whose own contribution amounts are always zero as free riders. Based on these classification criteria, 58.3% and 12.5% of subjects are classified as conditional cooperators and free riders, respectively, in the C-RM treatment. Likewise, 63.6% and 20.5% of subjects are classified as conditional cooperators and free riders, respectively, in the F-RM treatment.

The analysis shows that a substantial fraction of conditional cooperators did not disclose while some free riders did disclose their last-period action choices. The percentages of conditional cooperators and free riders who selected to costly disclose their past in the C-RM treatment are on average 25.8% and 21.1%, respectively. These two percentages in the F-RM treatment are 55.3% and 50.3%, respectively. It is intriguing that the disclosure rates are similar between conditional cooperators and free riders in the C-RM and F-RM treatments. The fit of the classification method based on Fischbacher *et al.* (2001) to subjects' contribution behavior can be examined by looking at subjects' contribution amounts in period 20 (the final period of the experiment). In period 20, a subject would not contribute positive amounts if he or she is purely selfish. Panels 1(b) and 2(b) of Figure 3 report the average contributions in period 20. These figures provide three interesting features, among others. First, as is consistent with Result 4, disclosers on average contribute higher amounts than non-disclosers even when only period 20 is considered. The ratios of the average contribution by disclosers to that by non-disclosers are around 2.9 and 8.2 in the C-RM and F-RM treatments, respectively (see the "All subjects" bars in the two panels). Second, disclosers on average contribute relatively large amounts even in period 20. The average contributions of disclosers are 9.2 ECUs and 7.4 ECUs in the C-RM and F-RM treatments, respectively. Third, and as importantly, the average contributions of free riders in period 20 are much lower than those of conditional cooperators. For instance, the average contributions of free riders in the C-RM treatment are 0.0 ECUs when disclosing their last-period actions (0.6 ECUs when not disclosing their last-period actions). The third observation implies

that the classified cooperation types are good indicators to measure subjects' contribution behavior.

Panels 1(a) and 2(a) of Figure 3 report average contribution amounts across all periods but periods 1 and 20 by disclosure decision. It indicates that both conditional cooperators and free riders contribute large amounts in the C-RM and F-RM treatments when they disclosed their last-period contribution amounts. The differences in the average contribution between conditional cooperators and free riders are not significant in both of the treatments (see columns (1) and (2) in Appendix Table C.3). This suggests that some free riders mimic the behavior of conditional cooperators by strategically contributing much in the C-RM and F-RM treatments.

RESULT 5: *(a) Regardless of disclosure decisions, free riders contribute much smaller amounts to the joint account in period 20, compared with conditional cooperators, in the C-RM and F-RM treatments. (b) However, free riders contribute almost similar amounts to conditional cooperators during the course of their plays (periods before period 20) when they disclose their last-period contribution amounts.*

In summary, the main reason behind the subjects' failure to sustain cooperation in the C-RM and F-RM treatments is that although disclosers, whether conditional cooperators or free riders, attempted to build cooperative reputations (Result 5), a substantial fraction of subjects did not disclose their past and then contributed only small amounts (Results 1 and 4).

4.3. Sorting Decisions and Action Choices in the C-Sorting and F-Sorting treatments

As discussed, subjects were successfully able to cooperate with each other in the reputation community than in the anonymous treatment. With a closer look, we found that although the level of cooperation was higher in the reputation community than in the anonymous community for the F-Sorting treatment, the cooperation norms in the reputation community was lower than that in the C-Sorting treatment. What drove the successful cooperation in the

reputation community? Why did the performance of the reputation community differ between the two treatments?

The C-Sorting and F-Sorting treatments have sorting mechanisms in that disclosers are matched with other disclosers in each period, unlike the C-RM and F-RM treatments. One possibility to answer the above two questions is that the composition of cooperation types are different between the reputation community and the anonymous community.[23] However, our data does not support this hypothesis. Appendix Figure C.3 reports the percentages of conditional cooperators and free riders in each of the reputation and anonymous communities. The data shows that these percentages are not significantly different between the reputation and anonymous communities in both the C-Sorting and F-Sorting treatments. In addition, as similar to Result 5(b), free riders contributed amounts similar to conditional cooperators in the reputation community.[24] The free riders' reputation building behavior is not surprising as they are assured to be paired with others disclosers thanks to the sorting devices embedded in the sorting treatments. These analyses suggest that the stronger cooperation norms emerged in the reputation community, especially in the C-Sorting treatment, were not due to the self-selection of cooperation-oriented subjects.

RESLT 6: *The percentages of conditional cooperators and free riders in the reputation community are not significantly different from those in the anonymous community in the C-Sorting and F-Sorting treatments. Moreover, the contribution behavior of conditional cooperators is not significantly different from that of free riders in the sorting treatments.*

[23] It has been demonstrated that subjects cooperate significantly more if they are sorted and are grouped with like-minded cooperative types in dilemma games (e.g., Gunnthorsdottir *et al.* 2007, Gächter and Thöni 2010).
[24] Conditional cooperators and free riders on average contributed 11.4 ECUs and 10.2 ECUs, respectively, to the joint accounts in the reputation community of the C-Sorting treatment; the difference in the average contribution is not statistically significant according to a session-level Mann-Whitney test (two-sided p-value = .2482). Conditional cooperators and free riders on average contributed 9.3 ECUs and 7.5 ECUs, respectively, to the joint accounts in the reputation community of the F-Sorting treatment; the difference in the average contribution is not statistically significant according to a session-level Mann-Whitney test (two-sided p-value = .3865).

The next possible factor that was behind the high efficiency in the reputation community is subjects' higher beliefs on their matched disclosers' contribution amounts.

First, our data shows that subjects' beliefs on their matched partners' contribution amounts significantly affect their own contribution behavior. This is true not only for the C-Sorting and F-Sorting treatments but also for the C-RM and F-RM treatments. Panels I(b) and II(b) in Figure 2 report the trends of the subjects' beliefs in the C-Sorting and F-Sorting treatments. The data revealed that disclosers' contribution amounts fluctuated parallel to their beliefs on matched partners' contributions (see red lines with diamonds in Panels (a) and (b)). The disclosers' average belief in the C-Sorting treatment was 12.15 ECUs, which was only .2 ECUs higher than the average of their own contribution amounts. Likewise, the disclosers' average belief in the F-Sorting treatment was 8.75 ECUs, which was only .28 ECUs lower than the average of their own contribution amount. The similar holds for non-disclosers (see blue line with triangles in Panels (a) and (b)).[25] A regression analysis confirms that contribution amounts and subjects' beliefs are significantly positively correlated for each of disclosers and non-disclosers in both the C-Sorting and F-Sorting treatments (columns (3) to (6) in Table 3(a)). The highly positive correlations between own actions and beliefs on the peers' action choices resonates with the idea that people experience psychological satisfaction from mutual cooperation (e.g., Fehr and Gächter 2000, 2002, Charness and Rabin 2002, Falk *et al.* 2005, Kamei and Putterman 2015, Rilling *et al.* 2002, Decety *et al.* 2004). The impact of subjects' beliefs on their own contributions is similarly observed in the C-RM and F-RM treatments (Table 3(a)).

Second, our data shows that subjects form beliefs based on their matched partners' last-period contribution amounts when the past contribution amounts are revealed to them. Table 3(b) reports a regression analysis in which the dependent variable is subject *i*'s belief on his or her

[25] For instance, the non-disclosers in the C-Sorting treatment on average believed that their partners would contribute 7.10 ECUs, which was only a little higher than the non-disclosers' average contribution, but it is not significantly different from the actual contribution amount.

period *t* matched partner's contribution amount and the independent variable includes the partner's reputation score. First, in the C-RM and F-RM treatments, subjects form significantly higher beliefs when they are matched with disclosers than when they are matched with non-disclosers. In addition, when subjects are matched with disclosers, their beliefs are positively correlated with the disclosers' last-period contribution amounts (Panel (b1)). Second, likewise, in the reputation community of the two sorting treatments, significantly positive correlations between disclosers' beliefs and period *t* matched partners' last-period contribution amounts are observed (Panel (b2)). Panel I(b) of Figure 2 shows that the much higher beliefs formed by disclosers in the C-Sorting treatment are almost parallel to the high contribution amounts their matched partners made in the last period which were informed to disclosers.

These two analyses suggest that a key for the success of high efficiency in the reputation community lies on the sorting mechanism where disclosers are matched with other disclosers. The impact of not having the sorting opportunity is substantial. Around 77% and 46% of the subjects did not disclose their last-period contribution amounts and then contributed only small amounts to the joint account in the C-RM and F-RM treatments, respectively (Result 4). The matched partners of the non-disclosers contributed significantly less in the C-RM and F-RM treatments as already discussed. This suggests an important role of the self-selection device in making rational cooperation successful.

RESULT 7: *(a) Subjects' contribution amounts and beliefs on their matched partners are significantly positively correlated in all of the four treatments. (b) The disclosers' beliefs and their matched partners' last-period contribution amounts are significantly positively correlated in the C-Sorting and F-Sorting treatments. (c) Subjects form significantly higher beliefs when they are matched with disclosers than otherwise; and the beliefs on the disclosers' contribution amounts are significantly positively correlated with the matched partners' last-period contribution amounts in the C-RM and F-RM treatments.*

As discussed earlier, Panels I(a) and II(a) of Figure 2 indicate a difference in the evolution of cooperation in the reputation community between the C-Sorting and F-Sorting treatments. Results 7(a) and (b) and Panels I(b) and II(b) of Figure 2 suggest that the difference in the level of cooperation between the two treatments is caused by the difference in the matched partners' last-period contribution amounts: the difference in partners' reputation score leads to the difference in beliefs on the partners' cooperativeness between the two treatments. Specifically, based on partners' previous behavior, disclosers formed much higher levels of beliefs in the C-Sorting treatment, whereas they formed only modest level of beliefs in the F-Sorting treatment. A likely reason for this difference is the free entrance to the reputation community in the F-Sorting treatment. In the literature of endogenous group formation, as discussed in Section 3, it is known that selfish individuals may chase cooperative ones attempting to exploit them and as a result cooperation may easily collapse (Ahn *et al.* 2008, 2009, Ehrhart and Keser 1999). The lack of the cost for sorting in the F-Sorting treatment may strengthen this negative effect of mobility. Figure 4 reports subject-by-subject average contribution amounts in the reputation community in the C-Sorting and F-Sorting treatments. This clearly shows that almost all subjects in the F-Sorting treatments very frequently joined the reputation community. The percentages of subjects who joined the reputation community more than ten times (except period 20), which we call the "frequent disclosers" hereafter, are 50.0% and 75.0% in the C-Sorting and F-Sorting treatment, respectively. The two percentages are significantly different (Two-sample test of proportions, two-sided p-value = .0154). Figure 4 also indicates that a larger percentage of the frequent disclosers contributed large amounts to the joint account in the C-Sorting treatment, compared with the F-Sorting treatment. For example, the percentages of the frequent disclosers that contributed amounts less than 10 ECUs are 9.1% and 40.9% in the C-Sorting and F-Sorting treatments, respectively. The small percentage in the C-Sorting treatment is striking. The two percentages are significantly different (Two-sample test of proportions, two-sided p-value = .0070). The percentages of the frequent disclosers that

contributed very small amount, less than 5 ECUs, are 2.2% and 15.9% in the C-Sorting and F-Sorting treatments, respectively; and the two percentages are weakly significantly different (Two-sample test of proportions, two-sided p-value = .0859). These analyses suggest that a significantly larger fraction of subjects with intentions to exploit high contributors joined the reputation community when joining is free, compared with when it is costly.

RESULT 8: *A significantly larger percentage of those who attempt to exploit high contributors join the reputation community in the F-Sorting treatment than in the C-Sorting treatment.*

Result 8 implies that the reputation community is more unstable in the F-Sorting treatment than in the C-Sorting treatment. This is because disclosers employ the discriminating strategies as summarized in Results 7(a) and (b); and some subjects' low contribution events would quickly spread to other subjects with the reputation system. This contagion process implied by Results 7 and 8 can explain the difference in the cooperation dynamics in the reputation community between the C-Sorting and F-Sorting treatments summarized in Result 3.

Lastly, we note that although the efficiency of the reputation community is significantly higher than that in the anonymous community especially in the C-Sorting treatment, a non-negligible fraction of the subjects stay away from the reputation community. This phenomenon cannot be explained by conditional cooperation types (Result 6). What accounts for this phenomenon? A regression analysis shows that this sorting outcome can be explained by the difference in people's expectation with regards to their peers' contribution behavior. In the regression, we used each subject's session-average beliefs on their matched peers' contribution amounts in the reputation community or anonymous community as the dependent variable. The independent variable includes the number of periods in which the subject joined the reputation community up to period 19. The data reveals that the reputation community in the C-Sorting treatment attracted subjects with more optimistic expectation about their peers' action choices. That is, the frequent members of the reputation community had significantly higher expectations

about their peers' contribution amounts than those who less frequently joined did in the same reputation community (column (1) in Table 5). By contrast, in the F-Sorting treatment, where the majority of the subjects are frequent disclosers, the beliefs formed by the frequent members of the reputation community are not significantly different from those by the infrequent members of it (column (3) in Table 5).

RESULT 9: *The frequent members of the reputation community have significantly higher expectations about their peers' contribution amounts in the reputation community than those who less frequently joined do so in the C-Sorting treatment.*

5. Conclusions

This study let subjects play a finitely-repeated two-person public goods game where each subject is given an option to disclose their past behavior. Our experiment showed that such information disclosure is not enough to sustain cooperation. A close look at the data reveals that the failure of cooperation with information disclosure is caused by (a) sufficiently low disclosure rates and the non-disclosers' opportunistic behavior and (b) subjects' discriminating strategies in that they contribute significantly less when faced with non-disclosers. However, it also showed that when subjects are given an opportunity to sort into the reputation community, a stable size of reputation community emerges and those who sort into the reputation community successfully cooperate with each other. Key causes for this stability are that (i) the reputation mechanism facilitates human cooperation by raising their expectation on others' pro-social acts, and (ii) the subjects adopt the 'tit-for-tat'-like conditional cooperation strategy based on their partners' past actions conveyed to them. We also found that the efficiency of such sorting opportunities diminishes if sorting into the reputation community is free because those who have malicious intentions to exploit high cooperators are more likely to join the reputation community, compared with when sorting is costly.

We remark that the emergence of anonymous communities in the sorting treatments fits observations in our real life as we see communities (e.g., online platforms) both with and without reputation mechanisms. Our detailed analysis suggests that a non-negligible fraction of conditional cooperators shy away from the reputation community due to their pessimistic beliefs on peers' action choices in the reputation community. What kind of additional institutions on top of the reputation mechanism would be helpful in reversing such subjects from the anonymous environment to the reputation environment? This question would be an interesting direction for future research.

As mentioned earlier, our paper has a useful contribution in the large body of the literature on rational cooperation in finitely-repeated dilemmas, on reputation mechanisms, and on endogenous group formation. We note that our paper also contributes to the literature on voluntary information disclosure which shows that disclosing private information of products or firms may raise the valuation of them. For instance, Lewis (2011) shows that on eBay motors, there is positive impact of voluntary information disclosure (e.g., photos) on the prices of used cars in auctions. Our experimental setup uses a simultaneous public goods game and it can describe a real-world exchange in which the interaction is taking place while both sides do not know the exact value they will get from the other. This kind of interaction has been increasingly popular in our life (e.g., the sharing economy such as Uber). This simultaneous setup is different from the above marketplace example where sellers move first and buyers move next. Our data shows that voluntary information disclosure alone may not be enough to lead to socially beneficial relationships in such simultaneous-move transactions; but if the information disclosure is linked to sorting devices, a substantial number of users may disclose their information and successfully enjoy the benefit of information disclosure within their reputation community.

Acknowledgement: This project was supported by research grants from the Telecommunications Advancement Foundation and from the Murata Science Foundation. The

author thanks Paudie Lynch (the IT manager at the Durham University Business School) for help in managing the ORSEE recruiting system when he conducted the experiment.

REFERENCES

Ahn, T.K., Mark Isaac, and Timothy Salmon, 2008. "Endogenous Group Formation." *Journal of Public Economic Theory* 10: 170-194.

Ahn, T.K., Mark Isaac, and Timothy Salmon, 2009. "Coming and Going: Experiments on Endogenous Group Sizes for Excludable Public Goods." *Journal of Public Economics* 93: 336-351.

Aimone, Jason, Laurence Iannaccone, Michael Makowsky, and Jared Rubin, 2013. "Endogenous Group Formation via Unproductive Costs," *Review of Economic Studies* 80: 1215-1236.

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

Andreoni, James, and John Miller, 1993. "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence." *Economic Journal* 103: 570-585.

Axelrod, Robert, 1984, *The evolution of cooperation*. New York: BasicBooks.

Axelrod, Robert, Hamilton William, 1981. "The Evolution of Cooperation." *Science* 211: 1390-1396.

Bayer, Ralph, 2011. "Cooperation in Partnerships: The Role of Breakups and Reputation." The University of Adelaide School of Economics Research Paper No. 2011-22.

Bolton Gary, Elena Katok, and Axel Ockenfels, 2004. "How Effective are Online Reputation Mechanisms? An Experimental Study." *Management Science* 50: 1587-1602.

Bolton Gary, Elena Katok, and Axel Ockenfels, 2005. "Cooperation among strangers with limited information about reputation." *Journal of Public Economics* 89: 1457-1468.

Camera, Gabriele, and Marco Casari, 2009. "Cooperation among Strangers under the Shadow of the Future." *American Economic Review* 99: 979-1005.

Charness, Gary, and Matthew Rabin, 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics* 117: 817-869.

Charness, Gary, and Yang Chun-Lei, 2014. "Starting Small toward Voluntary Formation of Efficient Large Groups in Public Goods Provision." *Journal of Economic Behavior & Organization* 102: 119-132.

Chaudhuri, Ananish, 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature." *Experimental Economics* 14: 47-83.

Cooper, Russell, Douglas DeJong, Robert Forsythe, and Thomas Ross, 1992. "Communication in Coordination Games." *Quarterly Journal of Economics* 107: 739-771.

Coricelli, Giorgio, Dietmar Fehr, and Gerlinde Fellner, 2004. "Partner Selection in Public Goods Experiments." *Journal of Conflict Resolution* 48: 356-378.

Cox, Caleb, Matthew Jones, Kevin Pflum, and Paul Healy, 2015. "Revealed reputations in the finitely repeated prisoners' dilemma." *Economic Theory* 58: 441-484.

Cox, James, Daniel Friedman, and Steven Gjerstad, 2007. "A tractable model of reciprocity and fairness." *Games and Economic Behavior* 59: 17-45.

Dal Bó, Ernesto, and Pedro Dal Bó, 2014. ""Do the right thing:" The effects of moral suasion on cooperation." *Journal of Public Economics* 117: 28-38.

Dal Bó, Pedro, Andrew Foster and Louis Putterman, 2014. "Institutions and Behavior: Experimental Evidence on the Effects of Democracy." *American Economic Review* 100: 2205-2229.

Decety, Jean, Philip Jackson, Jessica Sommerville, Thierry Chaminade, and Andrew Meltzoff, 2004. "The neural bases of cooperation and competition: an fMRI investigation." *NeuroImage* 23: 744-751.

Duffy, John, and Nick Feltovich, 2002. "Do Actions Speak Louder Than Words? Observation vs. Cheap Talk as Coordination Devices." *Games and Economic Behavior* 39: 1-27.

Duffy, John, and Nick Feltovich, 2006. "Words, Deeds, and Lies: Strategic Behaviour in Games with Multiple Signals," *Review of Economic Studies* 73: 669-688.

Dufwenberg, Martin, and Georg Kirchsteiger, 2004. "A theory of sequential reciprocity." *Games and Economic Behavior* 47: 268-298.

Ehrhart, Karl-Martin, amd Claudia Keser, 1999. "Mobility and Cooperation: on the Run." CIRANO working paper 99s-24.

Engelmann, Dirk, and Urs Fischbacher, 2009. "Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game." *Games and Economic Behavior* 67: 399-407.

Falk, Armin, Ernst Fehr, and Urs Fischbacher, 2005. "Driving Forces of Informal Sanctions." *Econometrica* 73: 2017-2030.

Fehr, Ernst, and Simon Gächter, 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90: 980-994.

Fehr, Ernst, and Simon Gächter, 2002. "Altruistic punishment in Humans." *Nature* 415: 137-140.

Fehr, Ernst, and Klaus Schmidt, 1999. "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics* 114: 817-868.

Fehr, Ernst, and Klaus Schmidt. 2006. ''The Economics of Fairness, Reciprocity and Altruism—Experimental Evidence and New Theories.'' In *Handbook of the Economics of Giving, Altruism and Reciprocity*, edited by S.-G. Kolm and J. M. Ythier, pp. 615-91. North Holland.

Fehr, Ernst, and Tony Williams, 2013. "Endogenous emergence of institutions to sustain," mimeo.

Fischbacher, Urs, 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10: 171-178.

Fischbacher, Urs, and Simon Gächter, 2010. "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments." *American Economic Review* 100: 541-56.

Fischbacher, Urs, Simon Gächter, and Ernest Fehr, 2001. "Are people conditionally cooperative? Evidence from a public goods experiment." *Economics Letters* 71: 397-404.

Gächter, Simon, and Christian Thöni, 2010. "Social Learning and Voluntary Cooperation among Like-Minded People." *Journal of European Economic Association* 3: 303-314.

Gallo, Edoardo, and Chang Yan, 2015. "The Effects of Reputational and Social Knowledge on Cooperation." *Proceedings of the National Academy of Science USA* 112: 3647-3652.

Gong, Binglin and Chun-Lei Yang, 2014. "Reputation and Cooperation: An Experiment on Prisoner's Dilemma with Second-order Information." Working paper.

Grimm, Veronika, and Friederike Mengel, 2009. "Cooperation in Viscous Populations – Experimental Evidence." *Games and Economic Behavior* 66: 202-220.

Gunnthorsdottir, Anna, Daniel Houser, and Kevin McCabe, 2007. "Disposition, history and contributions in public goods experiments." *Journal of Economic Behavior & Organization* 62: 304-315.

Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rockenbach, 2006. "The Competitive Advantage of Sanctioning Institutions." *Science* 312: 108-110.

Hauk, Esther, and Rosemarie Nagel, 2001. "Choice of Partners in Multiple Two-Person Prisoner's Dilemma Games." *Journal of Conflict Resolution* 45: 770-793.

Kamei, Kenju, 2015. "Endogenous Reputation Formation: Cooperation and Identity under the Shadow of the Future." Available at http://papers.ssrn.com/abstract=2556325.

Kamei, Kenju, 2016a. "Joint Decision-Making and Strategic Reputation Building in a Finitely-Repeated Dilemma." mimeo.

Kamei, Kenju, 2016b. "Cooperation and Endogenous Repetition in an Infinitely Repeated Social Dilemma: Experimental Evidence." mimeo.

Kamei, Kenju, and Louis Putterman, 2015. "Reputation Transmission without Benefit to the Reporter: a Behavioral Underpinning of Markets in Experimental Focus." Available at https://ideas.repec.org/p/bro/econwp/2015-9.html.

Kamei, Kenju, and Louis Putterman, forthcoming. "Play it Again: Partner Choice, Reputation Building and Learning from Finitely-Repeated Dilemma Games." *Economic Journal*.

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

Kreps, David, Paul Milgrom, John Roberts, and Robert Wilson, 1982. "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory* 27: 245-252.

Ledyard, John, 1995. "Public goods: A survey of experimental research." In J. H. Kagel and A.E. Roth (eds.), *The Handbook of Experimental Economics* 111-194, Princeton University Press.

Lewis, Gregory. 2011. "Asymmetric Information, Adverse Selection and Online Disclosure: The Case of eBay Motors." *American Economic Review* 101: 1535-1546.

Milinski, Manfred, Dirk Semmann, and Hans-Jürgen Krambeck, 2002. "Reputation helps solve the 'tragedy of the commons'." *Nature* 415: 424-426.

Nicklisch, Andreas, Kristoffel Grechenig, and Christian Thöni, 2015. "Information-sensitive Leviathans – the Emergence of Centralized Punishment." WiSo-HH Working Paper No. 13.

Nowak, Martin, Karen Page, and Karl Sigmund, 2000. "Fairness versus reason in the ultimatum game." *Science* 289: 1773-1775.

Nowak, Martin, and Karl Sigmund, 1998a. "Evolution of Indirect Reciprocity by Image Scoring." *Nature* 393: 573-577.

Nowak, Martin, and Karl Sigmund, 1998b. "The dynamics of indirect reciprocity." *Journal of Theoretical Biology* 194: 561-574.

Page, Talbot, Louis Putterman, and Bulent Unel, 2005. "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency." *Economic Journal* 115: 1032-1053.

Rabin, Matthew, 1993. "Incorporating Fairness Into Game Theory and Economics." *American Economic Review* 83: 1281-1302.

Reuben, Ernesto, and Sigrid Suetens, 2012. "Revisiting Strategic versus Non-strategic Cooperation." *Experimental Economics* 15: 24-43.

Rilling, James, David Gutman, Thorsten Zeh, Giuseppe Pagnoni, Gregory Berns, and Clinton Kilts, 2002. "A Neural Basis For Social Cooperation." *Neuron* 35: 395-405.

Rockenbach, Bettina, and Manfred Milinski, 2006. "The efficient interaction of indirect reciprocity and costly punishment." *Nature* 444: 718-723.

Seinen, Ingrid, and Arthur Schram, 2006. "Social status and group norms: Indirect reciprocity in a helping experiment." *European Economic Review* 50: 581-602.

Selten, Reinhard, and Rolf Stoecker, 1986. "End Behaviour in sequences of finite prisoner's dilemma supergames: a learning theory approach." *Journal of Economic Behavior and Organization* 7: 47-70.

Sobel, Joel, 2005. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature* 43: 392-436.

Tyran, Jean-Robert, and Lars Feld, 2006. "Achieving Compliance when Legal Sanctions are Non-Deterrent." *Scandinavian Journal of Economics* 108: 135-56.

Wedekind, Claus, and Victoria Braithwaite, 2002. "The Long-Term Benefits of Human Generosity in Indirect Reciprocity." *Current Biology* 12: 1012-15.

Wedekind, Claus, and Manfred Milinski, 1996. “Human Cooperation in the Simultaneous and the Alternating Prisoner’s Dilemma: Pavlov Versus Generous Tit-For-Tat.” *Proceedings of the National Academy of Sciences USA* 93: 2686- 2689.

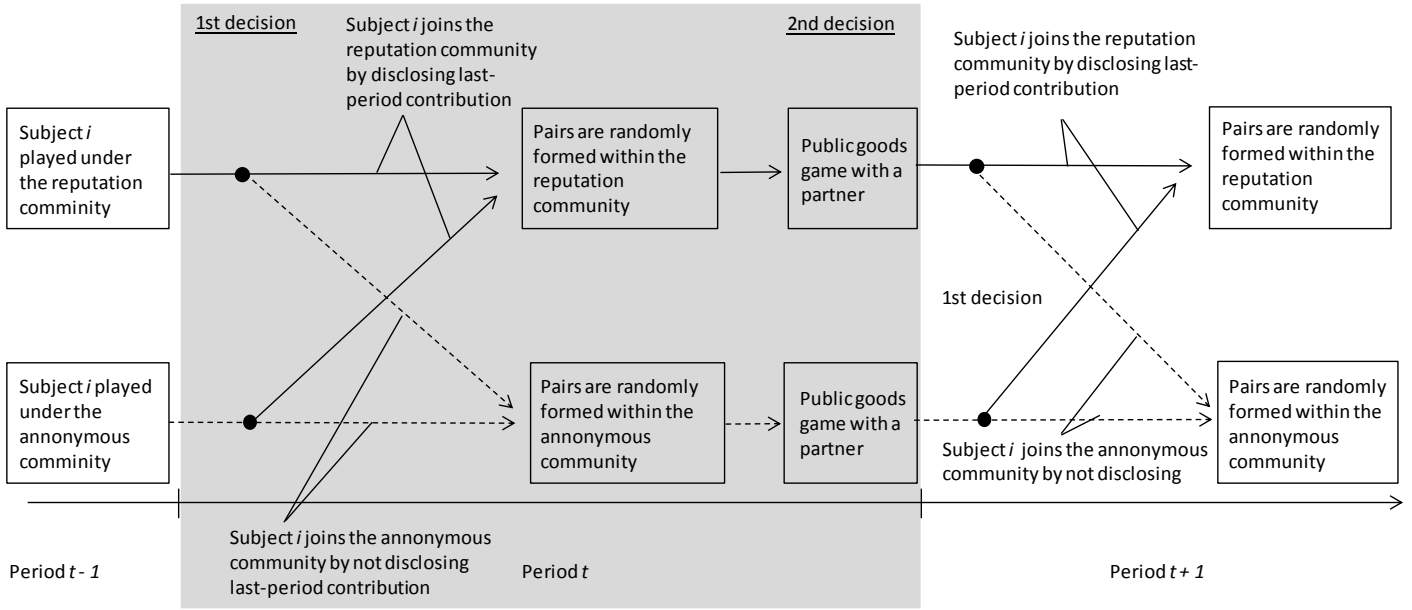
Wedekind, Claus, and Manfred Milinski, 2000. “Cooperation Through Image Scoring in Humans.” *Science* 288: 850-852.

Table 1. *Summary of Treatments*

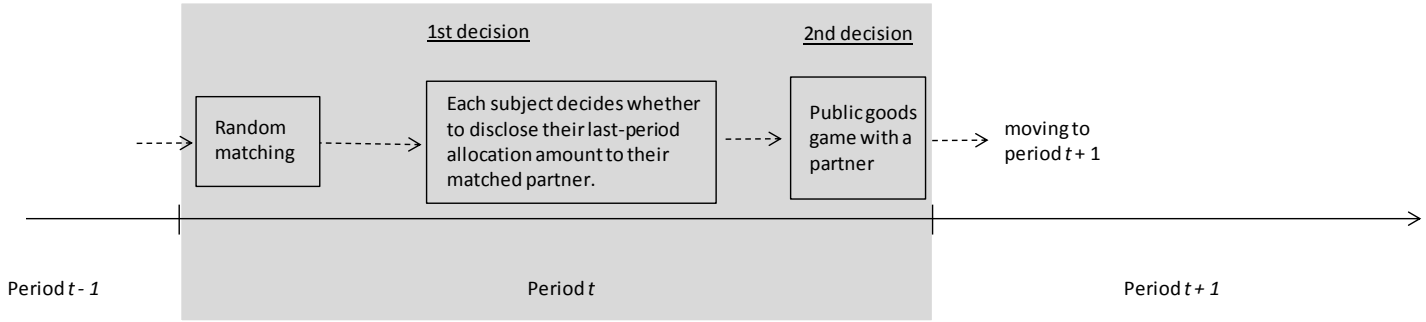
Treatment name	Costs for disclosing/sorting into the reputation community	Matching protocol	The number of sessions (subjects)	The fraction of those who disclosed or sorted into the reputation community	Average contribution		
					All data	reputation community	anonymous community
C-Sorting	1 ECU	Sorting: Each discloser (non-discloser) is matched with another discloser (non-discloser).	4 (44)	52.8%	9.28 (46.3%)	11.96 (59.8%)	6.15 (30.7%)
C-RM	1 ECU	Random Matching: Each subject is randomly matched with another subject.	4 (48)	23.0%	6.02 (30.1%)	----	----
F-Sorting	0 ECUs	Sorting: Each discloser (non-discloser) is matched with another discloser (non-discloser).	4 (44)	70.7 %	7.89 (39.4%)	9.03 (45.2%)	5.93 (29.7%)
F-RM	0 ECUs	Random Matching: Each subject is randomly matched with another subject.	4 (44)	53.9%	7.62 (38.1%)	----	----

Notes: C-Sorting = Costly Sorting. F-Sorting = Free Sorting. C-RM = Costly Disclosure, Random Matching. F-RM = Free Disclosure, Random Matching.

Figure 1. *Schematic Diagram*



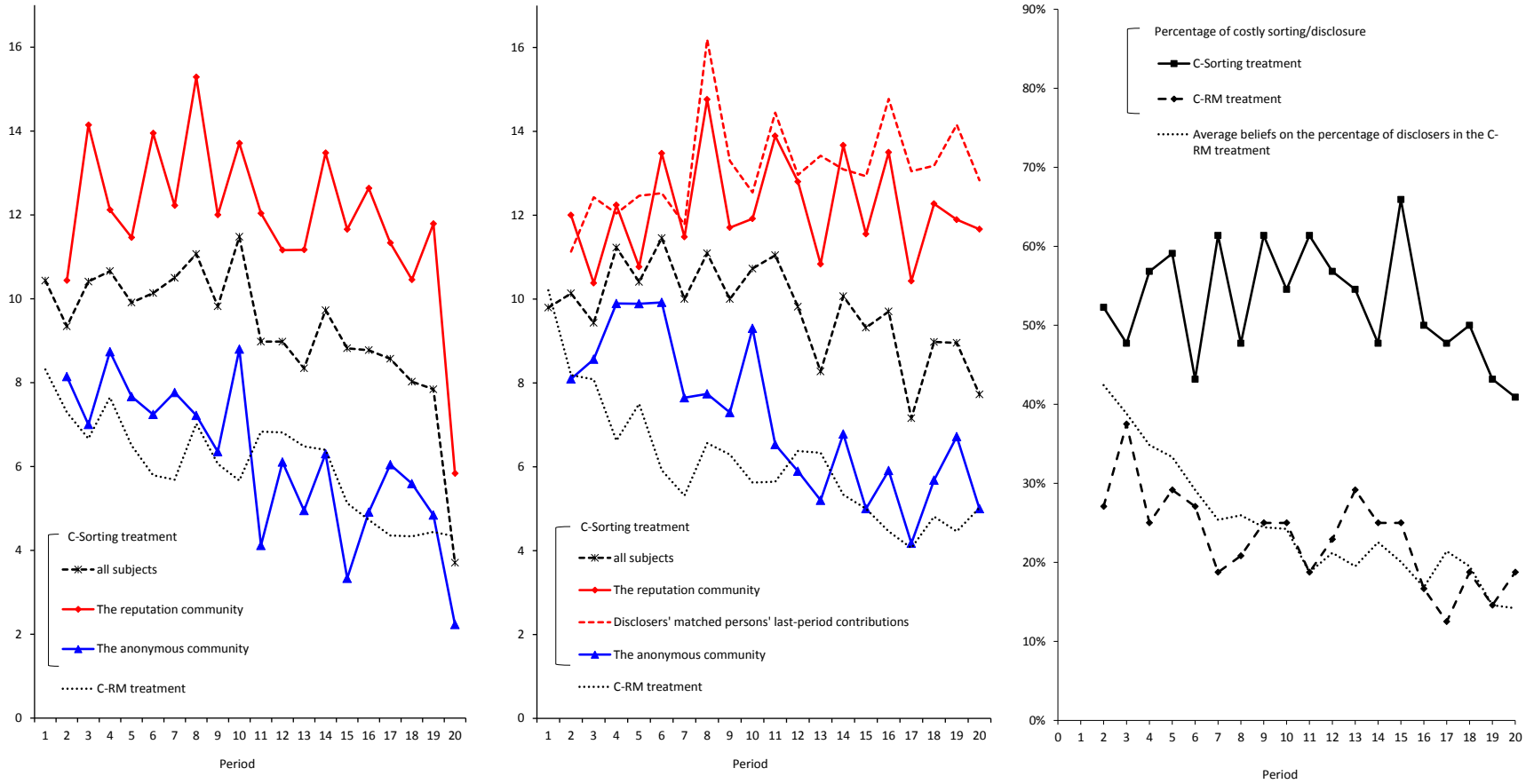
(a) C-Sorting and F-Sorting treatments



(b) C-RM and F-RM treatments

Notes: At the onset of the experiment, subjects go through the task to elicit conditional cooperation types. There is no disclosure decision stage in period 1. In the allocation stage, subjects in all treatments are asked to state beliefs on their matched partner's contribution amount in a given period. In the C-RM and F-RM treatments, subjects are also asked to answer guess on how many persons disclosed immediately after their disclosure decision in a given period.

Figure 2. *Period-by-Period Average Contributions and the Percentage of the Subjects Who Joined the Reputation Community*

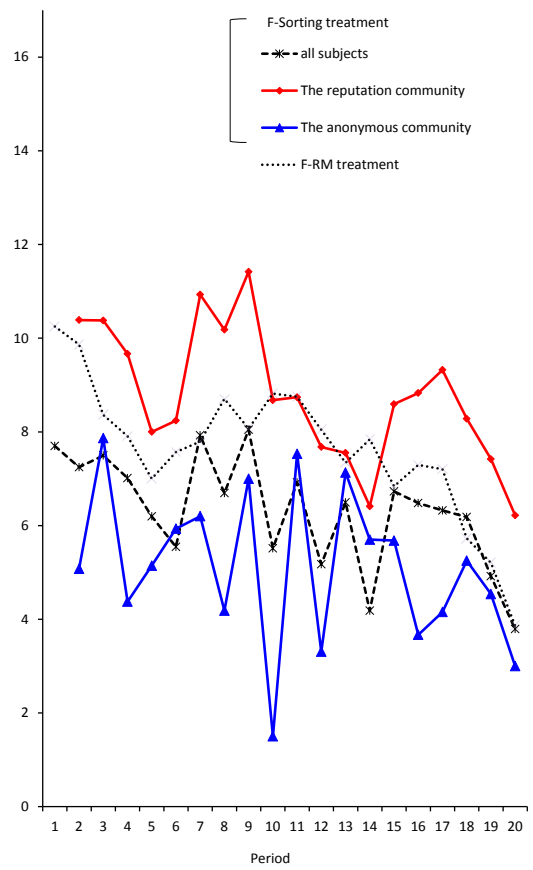


(a) Average contributions

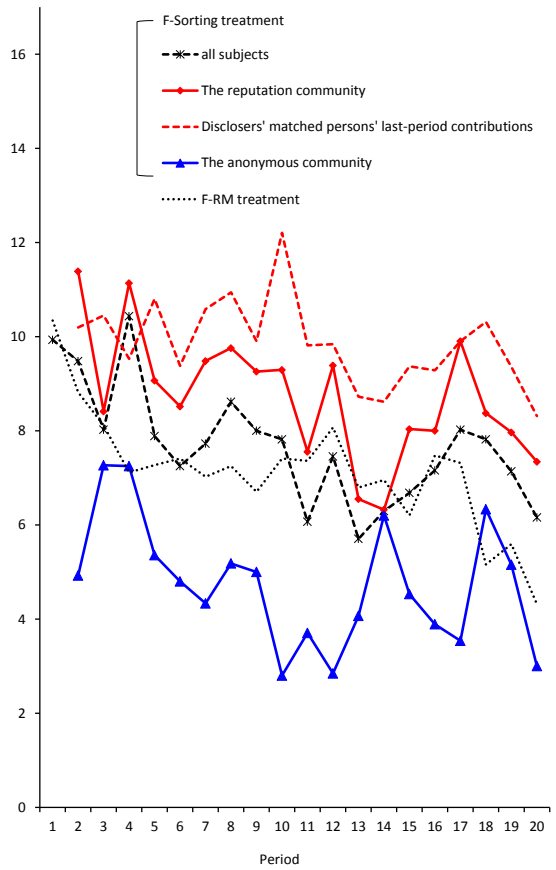
(b) Average beliefs on their matched partners' contributions

(c) The percentage of those who disclosed/ sorted into the reputation community

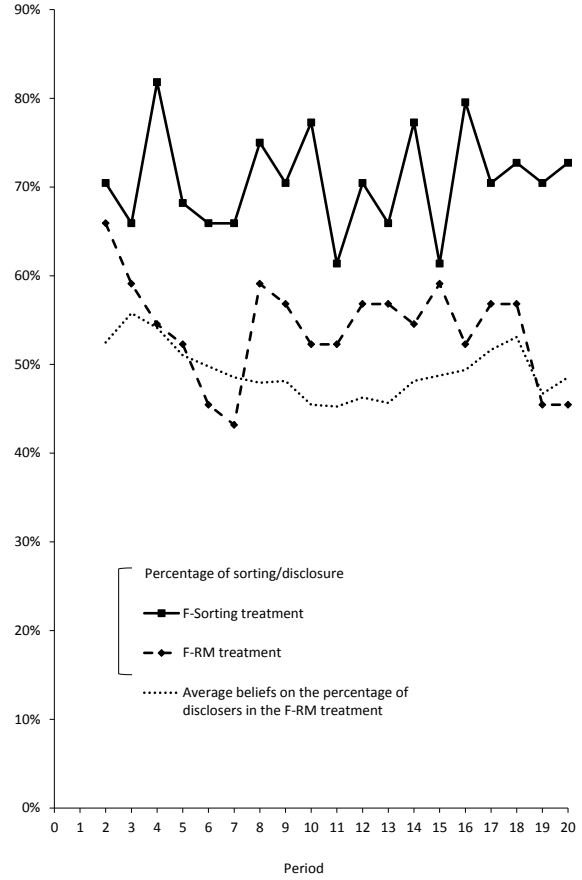
(I) C-Sorting and C-RM treatments



(a) Average contributions



(b) Average beliefs on their matched partners' contributions



(c) The percentage of those who disclosed/ sorted into the reputation community

(II) F-Sorting and F-RM treatments

Notes: The red dash line depicted in figure (b) shows the average of last-period contributions made by the disclosers' matched partners to the joint accounts in the C-Sorting (F-Sorting) treatment in Panel (I) (Panel (II)). See Appendix Figure C.1 for period-by-period diagrams of subjects' moving-in and moving-out of the reputation community in the C-Sorting and F-Sorting treatments, which shows that there are a stable number of events with moving-in and moving-out in every period in the two sorting treatments.

Table 2. *Average Contribution Trends by Treatment*

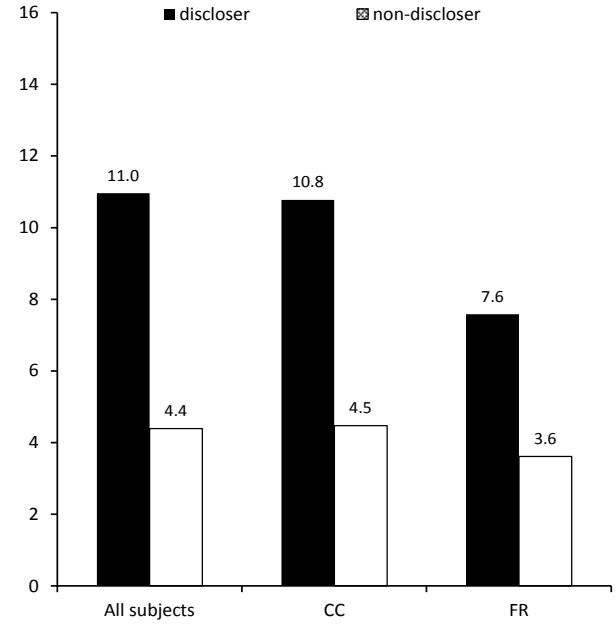
Dependent variable: Session-average contribution amounts to their joint account in period *t*

Treatment: / Independent Variable:	C-RM	C-Sorting		F-RM	F-Sorting	
		The reputation community[#2]	The anonymous community[#3]		The reputation community[#2]	The anonymous community[#3]
	(1)	(2)	(3)	(4)	(5)	(6)
Period Number[#1]	-.16***	-0.049	-0.26***	-.16***	-.15**	-.019
	(.031)	(0.069)	(0.058)	(.046)	(.059)	(.086)
Constant	7.76***	12.6***	9.43***	9.40***	10.9***	5.10***
	(.35)	(0.81)	(0.68)	(.52)	(.69)	(1.00)
# of Observations	76	72	72	76	72	70[#4]
F	28.26	0.51	20.55	11.81	6.57	.05
Prob > F	.0000	.4796	.0000	.0010	.0126	.8252
R-squared	.1847	.0065	.2169	.0998	.0499	.0005

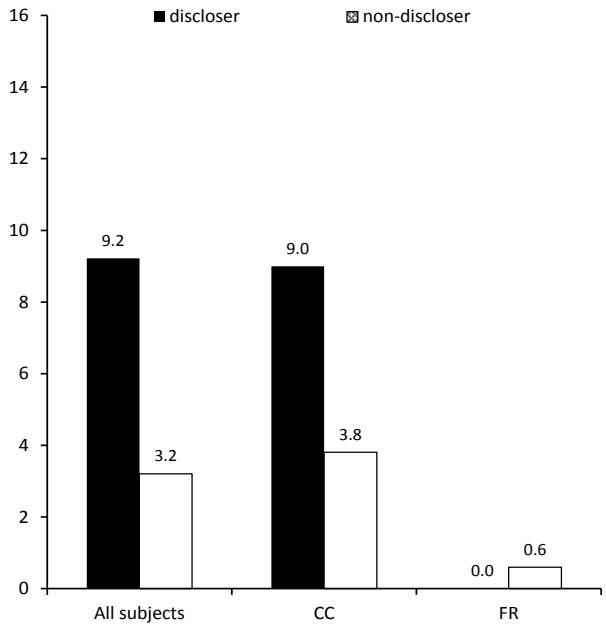
Notes: Session fixed-effects linear regressions. The numbers in parentheses are standard errors. [#1] The Period Number variable equals 2, 3, …, or 19 for the C-Sorting and F-Sorting treatments. Observations in period 1 were not used as there were no self-selection decisions in that period. The Period Number variable equals 1, 2, …, 19 for the C-RM and F-RM treatments. Observations in period 20 were not included in all of the four treatments because the usual end-game defection was observed (Andreoni 1988). [#2] The reputation community = the set of disclosers. [#3] The anonymous community = the set of non-disclosers. [#4] All subjects chose to sort into the reputation community in periods 15 and 16 in one session. The number of observations is therefore 70 unlike the reputation community in column (5).

*, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

Figure 3. *Average Contribution Amounts by Disclosure Decision in the C-RM and F-RM treatments*

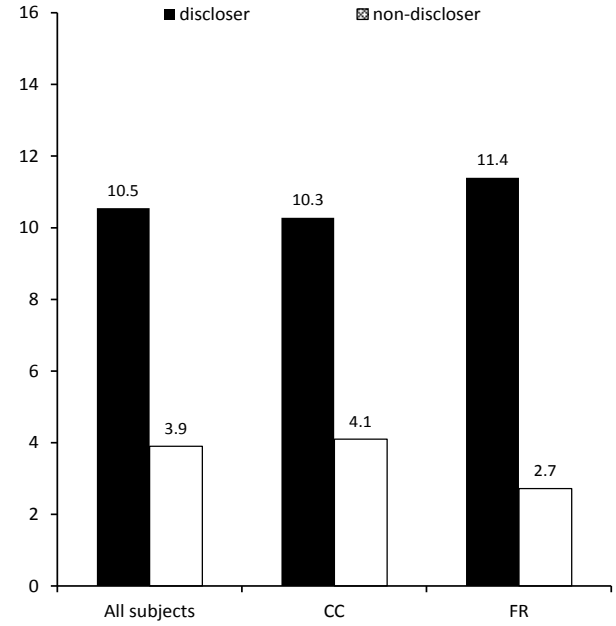


(a) Average across all periods

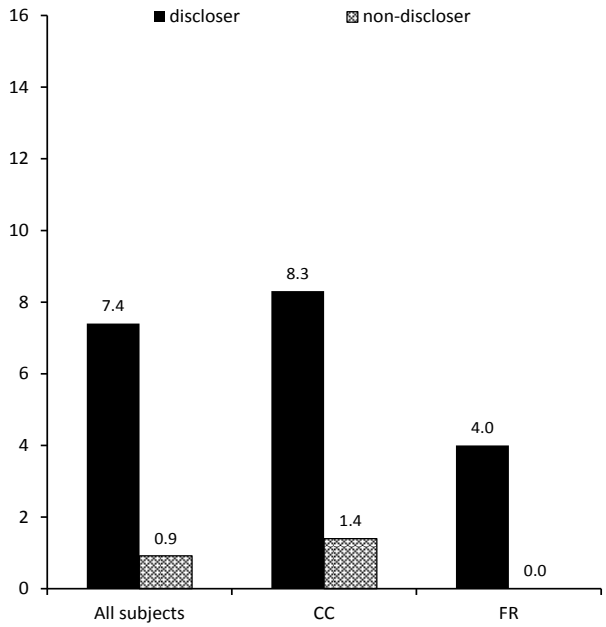


(b) Period 20 (final period of the experiment)

(1) The C-RM treatment



(a) Average across all periods



(b) Period 20 (final period of the experiment)

(2) The F-RM treatment

Notes: CC and FR refer to conditional cooperators and free riders, respectively. The 'all subjects' category includes also cooperation types other than conditional cooperators and free riders.

Table 3. *Subjects' Contributions, Beliefs, and Partners' Last-Period Contribution Amounts*

(a) Relationship between the Subjects' Contributions and Beliefs on their Partners' Contribution Amounts

Dependent variable: Subject *i*'s contribution amount to his or her joint account in period *t*, where $t \in \{2, 3, \ldots, 19\}$.

Data:	The random-matching community		The reputation community (the set of disclosers)[#2]		The anonymous community (the set of non-disclosers)[#3]	
	(1)[#1]	(2)[#1]	(3)	(4)	(5)	(6)
Independent Variable:	C-RM	F-RM	C-Sorting	F-Sorting	C-Sorting	F-Sorting
(i) Subject *i*'s belief on his or her matched partner's contribution amount in period *t*	.14*** (.0079)	.12*** (.019)	.11*** (.015)	.12*** (.018)	.12*** (.015)	.098* (.054)
# of Observations	864	792	423	559	369	233
Log pseudolikelihood	-1433.0	-1467.7	-730.3	-1037.0	-779.0	-355.9
Wald Chi-squared	307.3	41.02	57.36	47.29	57.45	3.27
Prob > Wald Chi-squared	.0000	.0000	.0000	.0000	.0000	.0706

Notes: Individual random-effects ordered probit regressions with standard errors clustered by session. Estimates of cut points are omitted to conserve space. Observations in period 20 were not included because the usual end-game defection was observed (Andreoni 1988). The results are similar even if observations in period 20 are included (the results are omitted in the table to conserve space).

[#1] We also ran the same regressions while also having a dummy variable which equals 1 (0) if subject *i*'s period *t* partner did not disclose (disclosed) his or her period $t-1$ contribution amount as an independent variable. The analysis shows that the dummy variable fails to obtain a significant coefficient while independent variable (i) obtains a significantly positive coefficient (the size of the coefficient estimates are almost similar to the ones in the above table). The results are omitted to conserve space.

[#2] A small number of the disclosers were matched with non-disclosers as explained in the text. Results in columns (3) and (4) are almost similar even if we only use the observations in which discloser *i* was matched with another discloser *j*. The results are omitted to conserve space.

[#3] A small number of the non-disclosers were matched with disclosers as explained in the text. Results in column (5) are almost similar even if we only use the observations in which non-discloser *k* was matched with another non-discloser *m*. As for results in column (6), if we only use the subset of data, the coefficient of variable (i) obtains a statistical significance at the 1% level while the size of the coefficient estimate is similar to the one in column (6). The results are omitted to conserve space.

*, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

(b) Relationship between the Subjects' Beliefs on their Partners' Contribution Amounts and the Partners' Last-period Contributions

(b1) The C-RM and F-RM treatments

Dependent variable: Subject *i*'s belief on his or her matched person *j*'s contribution amount in period *t*, where $t \in \{2, 3, \ldots, 19\}$.

Independent Variable:	C-RM (1)	F-RM (2)
(1 – the No Information dummy) × *j*'s period *t* – 1 contribution amount	.061*** (.0018)	.076*** (.017)
# of Observations	864	792
Log pseudolikelihood	-1701.1	-1566.2
Wald Chi-squared	1108.9	19.34
Prob > Wald Chi-squared	.0000	.0000

Notes: Individual random-effects ordered probit regression with standard errors clustered by session. The reference group is those whose matched partners did not disclose their last-period contribution amounts. Observations in period 20 were not included because of the strong end-game defection observed in the experiment. However, results are similar even if observations in period 20 was included (the results are omitted to conserve space). The No Information dummy equals 1 if subject *i*'s period *t* matched person *j* did not disclose his or her last-period contribution amount in period *t*; and 0 otherwise. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

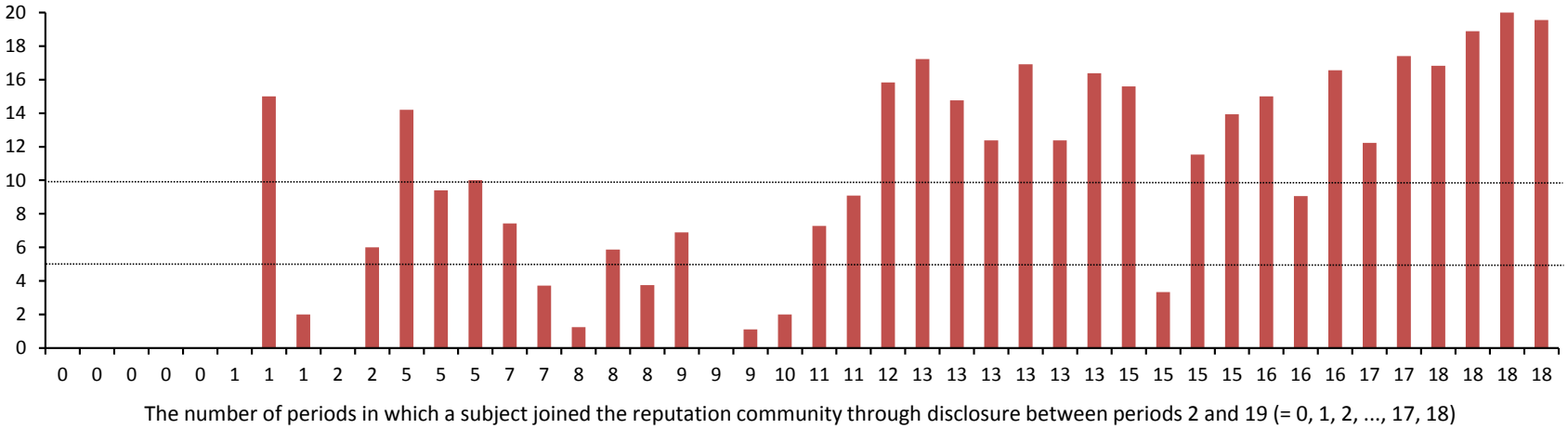
(b2) The Reputation Community in the C-Sorting and F-Sorting treatments

Dependent variable: Subject *i*'s belief on his or her matched discloser *j*'s contribution amount in period *t*, where $t \in \{2, 3, \ldots, 19\}$, in the reputation community.

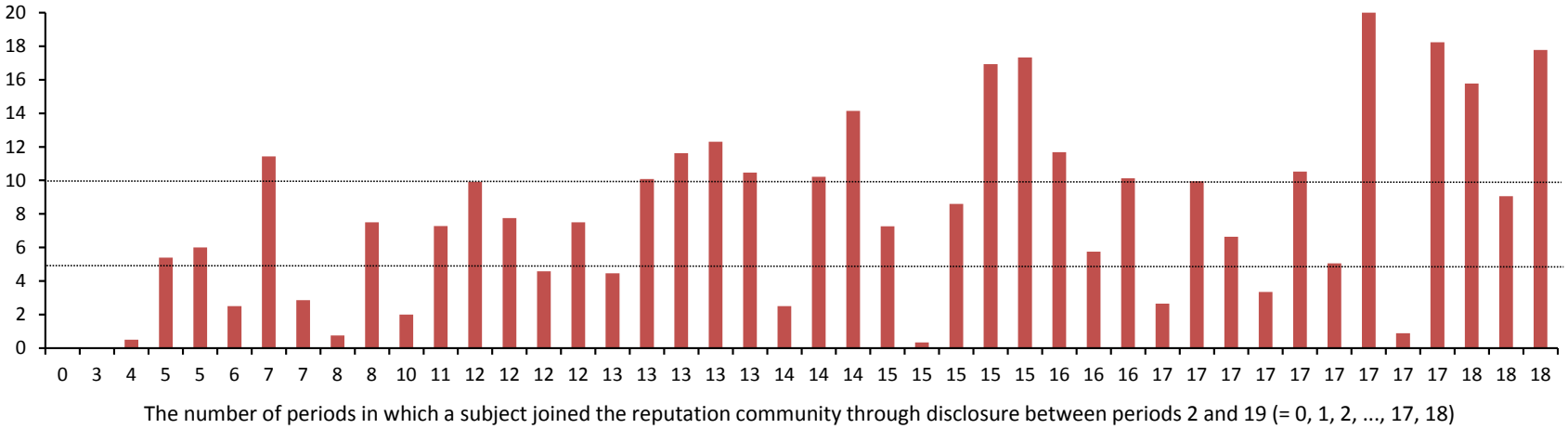
Independent Variable:	C-Sorting (1)	F-Sorting (2)
Subject *i*'s period *t* matched discloser *j*'s period *t* – 1 contribution amount	.11*** (.026)	.079*** (.010)
# of Observations	384	520
Log pseudolikelihood	-690.2	-1028.7
Wald Chi-squared	18.71	59.75
Prob > Wald Chi-squared	.0000	.0000

Notes: Individual random-effects ordered probit regression with standard errors clustered by session. Only observations in which discloser *i* was matched with another discloser *j* were used in the regression analyses. Observations in period 20 were not included because of the strong end-game defection observed in the experiment. However, results are similar even if observations in period 20 was included (the results are omitted to conserve space). *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

Figure 4. *Average Contribution Amounts of Disclosers in the Reputation Community by the Number of Periods in Which Subjects Joined the Reputation Community*



(1) The C-Sorting treatment



(2) The F-Sorting treatment

Note: Each bar indicates each subject's average contribution when he or she joined the reputation community.

Table 5. *Beliefs on the Matched Partners' Contribution Amounts and the Frequencies of Joining the Reputation Community in the C-Sorting and F-Sorting treatments.*

Dependent variable: Subject *i*'s session average belief on his or her partners' contribution amounts to the joint accounts when *i* joined the reputation community (in columns (1) and (3)) or the anonymous community (in columns (2) and (4)) up to period 19

	C-Sorting treatment		F-Sorting treatment	
Independent Variable:	Reputation community (1)	Anonymous community (2)	Reputation community (3)	Anonymous community (4)
The total number of periods in which subject *i* joined the reputation community up to period 19	.37**	.084	.17	.041
	(.15)	(.14)	(.16)	(.19)
Constant	7.17***	6.81***	6.37***	4.44*
	(1.76)	(1.48)	(2.24)	(2.50)
# of Observations	39[#1]	40[#2]	43[#3]	41[#4]
F	6.53	.35	1.09	.05
Prob > F	.0148**	.5551	.3028	.8306
R-squared	.1271	-.0168	.0021	-.0244

Notes: Linear regressions.

[#1] Five subjects had never joined the reputation community until period 19 in the C-Sorting treatment; thus, the total number of observations is 39 (= 44 – 5).

[#2] Four subjects had always joined the reputation community until period 19 in the C-Sorting treatment; thus, the total number of observations is 40 (= 44 – 4).

[#3] One subject had never joined the reputation community until period 19 in the F-Sorting treatment; thus, the total number of observations is 43 (= 44 – 1).

[#4] Three subjects had always joined the reputation community until period 19 in the F-Sorting treatment; thus, the total number of observations is 41 (= 44 – 3).

*, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.