Article

RATIONALITY AND SOCIETY

Rationality and Society
2016, Vol. 29(4) 471–503

Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1043463117734171
journals.sagepub.com/home/rss
SAGE

Boosting trust by facilitating communication: A model of trustee investments in information sharing

Vincenz Frey
Utrecht University, The Netherlands

Abstract

Trust problems hamper many social and economic exchanges. In such situations, there are often institutions that enable trustors to share information on the performance of trustees. While the benefits of such institutions have been researched extensively, little is known about their emergence. This article presents a game-theoretic model for the understanding of investments by trustees in establishing information sharing between trustors. The model allows for a simultaneous analysis of investments in and effects of institutions for information sharing. It captures two mechanisms by which a trustee's investment can promote trust. First, a trustee's investment in establishing information sharing can enable *network effects* that facilitate trust and trustworthiness. Second, it can promote trust by serving as a *signal* of intrinsic trustworthiness. The analysis of the model implies predictions for how characteristics of the interaction situation affect whether these mechanisms motivate a trustee to establish information sharing.

Keywords

Network effects, network formation, reputation, signaling, social dilemma, trust

Corresponding author:
Vincenz Frey, Department of Sociology, Interuniversity Center for Social Science Theory and Methodology (ICS), Utrecht University, Padualaan 14, Room C2.02, 3584 CH Utrecht, The Netherlands.
Email: v.c.frey@uu.nl

Introduction

Exchange often requires that one party—the ''trustor''—exposes herself to the risk of abuse by the other party—the ''trustee'' (Coleman, 1990; Dasgupta, 1988; Kreps, 1990). Buying a used car is a well-known example. The buyer (trustor) does not know the history of the car and the seller (trustee) may benefit from concealing important information. If the buyer places trust and buys, she risks paying too much. Trust problems are also salient in online trade: the buyer typically has to pay upfront without having any real guarantee that the seller will ever ship the product. Such trust problems pose a social dilemma. Both parties would benefit from fair exchange, but the exchange may not take place because the trustor has reason to mistrust the trustee (Akerlof, 1970; Coleman, 1990; Kollock, 1998). Fortunately, institutions that enable trustors to share information about trustees—as, for example, word-of-mouth networks or online reputation systems—allow actors to escape the dilemma (Buskens and Raub, 2002, 2013; Cook et al., 2009; Feinberg et al., 2014; Milinski, 2016; Resnick et al., 2000).

While it is well understood how institutions for information sharing facilitate exchange, little is known about the emergence of such institutions. There is a general theoretical idea that if certain institutions or networks have value for actors to achieve their goals, actors should be likely to set up such institutions or networks (Flap, 2004; Lin, 2002: Chap. 8; Prendergast, 1999). It also suggests itself that many institutions for information sharing have been set up purposively to facilitate exchange. In line with this idea, Guseva and Rona-Tas (2001) suggest that trust problems in credit markets led to the installment of credit bureaus in the United States and that these trust problems led Russian bankers to form extended word-of-mouth networks (see Klein, 1997 for related case studies). Still, there is a lack of theoretical models for the understanding of investments in establishing institutions for information sharing as a means to facilitate trust and trustworthiness. This article addresses this gap in the literature. It investigates theoretically under what circumstances it is likely that a trustee makes a costly investment to enable information sharing between trustors. By treating institutions for information sharing as endogenous, this article furthermore brings to attention effects of such institutions that have previously been overlooked. The article thus contributes to the understanding of the emergence *and* effects of institutions for information sharing.

The game-theoretic model presented in this article captures two mechanisms by which a trustee's investment in establishing information sharing can promote exchange. First, there are the well-known *network*

effects. Information sharing facilitates exchange by enabling trustors to learn about trustees from the experiences of other trustors. In addition, it gives trustees incentives to act honestly because developing a bad reputation could inhibit many future exchanges or require a trustee to lower his price.[1] These network effects are well-established theoretically as well as empirically (Abraham et al., 2016; Bohnet et al., 2005; Bohnet and Huck, 2004; Bolton et al., 2004; Buskens et al., 2010; Buskens and Raub, 2002, 2013; Cook and Hardin, 2001; DiMaggio and Louch, 1998; Feinberg et al., 2014; Frey and Van de Rijt, 2016; Hillmann and Aven, 2011; Huck et al., 2010; Przepiorka, 2013; Snijders and Weesie, 2009; Sosis, 2005). In the current study, I model how these network effects can motivate a trustee to set up an institution for information sharing between trustors.

Second, the model shows that a trustee's investment in establishing information sharing can promote trust by serving as a *signal* of intrinsic trustworthiness. Some people are intrinsically trustworthy and behave well also in the absence of contextual factors that deter trust abuse.[2] But how could a trustee convince a trustor that he is intrinsically trustworthy given that also those with bad intentions want to be trusted? Signaling theory (e.g. Bliege Bird and Smith, 2005; Gambetta, 2009; Przepiorka and Berger, 2017; Spence, 1973; Zahavi, 1975) maintains that taking some costly, observable action can allow an actor to credibly communicate unobservable characteristics, such as intrinsic trustworthiness (Bacharach and Gambetta, 2001; Barclay, 2004; Bliege Bird and Power, 2015; Gambetta and Przepiorka, 2014; Paik and Woodley, 2012; Przepiorka and Berger, 2017; Przepiorka and Diekmann, 2013; Raub, 2004). This article shows that a trustee's investment in establishing information sharing can serve as such a signal of intrinsic trustworthiness.

I model these network and signaling effects and how they can motivate a trustee to invest in establishing information sharing using the framework of finitely repeated Trust Games (TGs) with incomplete information (see, e.g. Buskens et al., 2017; James, 2002). The model focuses on a scenario in which several trustors interact with the same trustee. The trustors do not know whether the trustee is of the *friendly type* or of the *opportunistic type*. A trustee of the friendly type is intrinsically trustworthy and has no incentive to abuse trust. A trustee of the opportunistic type has an incentive to abuse trust, at least in the short run. At the beginning of the game, the trustee can make a costly investment to set up an institution for information sharing *between the trustors*. I identify under what conditions there exist (sequential) equilibria in which the trustee pledges this investment to allow the network effects or to signal that he is of the friendly type.

To my knowledge, Frey et al. (2015) and Raub et al. (2013) are the only studies that investigate how network effects can motivate actors to establish information sharing. The paper by Frey et al. (2015) is a companion paper to the current study. It investigates in a very similar game-theoretic model investments in information sharing by *trustors* rather than trustees. Raub et al. (2013) study a game-theoretic model that allows also investments by trustees, but their model assumes complete information about the incentives of trustees and indefinite repetition. A core result of Frey et al. (2015) and Raub et al. (2013) is that there is an inverse U-shape in the relation between the incentives to invest in establishing information sharing and the size of the trust problem. The incentives to invest are small for trust problems that are very small or very large, while the incentives to invest are large for trust problems of intermediate size. The results of this article regarding investments motivated by the network effects are in line with this prediction.

That a trustee's investment in establishing information sharing can serve as a signal of intrinsic trustworthiness has not been studied previously. This innovation leads to new predictions on investments in information sharing as well as effects of information sharing. Furthermore, the article contributes to the broader literature on signaling trustworthiness (see Przepiorka and Berger, 2017 for a recent overview). I analyze the model for two types of friendly trustees: friendly trustees who would suffer from an *internal sanction* if abusing trust and friendly trustees who are trustworthy because they receive an *internal reward* if honoring trust. These scenarios subsume various motivations for intrinsic trustworthiness, such as inequity aversion, guilt aversion, altruism, or warm-glow.[3] We will see that the possibility to signal intrinsic trustworthiness depends crucially on what motivation leads to intrinsic trustworthiness and that results concerning signaling equilibria generalize beyond the specific context of investments in information sharing.

The game

This section describes the game Γ and the difference between the two versions that I consider—the version in which friendly trustees are *g*uilt-*a*voiding trustees (Γ^{ga}) and the version in which friendly trustees are *r*eward-*s*eeking trustees (Γ^{rs}). Throughout, I use the superscripts *ga* and *rs* to let notation refer specifically to either of these versions of Γ. I omit these superscripts if I refer to both versions simultaneously.

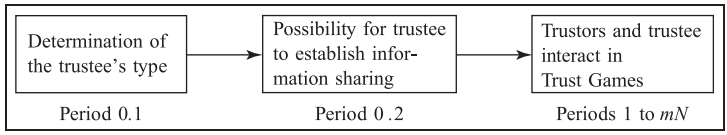

Figure 1. Timeline of Γ.

Actors, moves, and information

Γ has two types of actors: $m \geq 2$ trustors and one trustee. It proceeds as illustrated in Figure 1. First, the trustee's type is determined and the trustee learns his type. Then, the trustee decides whether to invest in establishing information sharing. Finally, the trustors and the trustee interact in TGs.

More specifically, in period 0.1, a random move of a pseudo-actor Nature determines the type of the trustee. With probability π the trustee is of the *friendly type* (the intrinsically trustworthy type) and with probability $1 - \pi$ he is of the *opportunistic type*, where $0 < \pi < 1$. Friendly and opportunistic trustees differ in their payoffs in the TG as described below. The probability π is common knowledge (each actor knows π, knows that the others know that (s)he knows π, and so on). The trustee also knows his actual type, but the trustors are not informed about the trustee's type.

In period 0.2, the trustee decides whether to invest the cost c to establish an institution for information sharing *between the m trustors*.

Finally, one binary TG is played in each of the periods 1 to mN. In every TG, the trustor at play, first, decides whether to place trust. If the trustor does not place trust, the TG ends. If the trustor places trust, the trustee chooses whether to honor or abuse trust and the TG ends.

The order in which the trustors play their TGs is random but each of the m trustors plays precisely N TGs with the trustee. The analysis does not require that the order of interaction is known in advance. It suffices to assume perfect recall (i.e. that players never forget information they once acquired), so that actors remember information on outcomes in specific periods and who played in which period.

If the trustee established information sharing in period 0.2, all trustors are always informed about the choices that were made in *all* past TGs. Information is shared truthfully among the m trustors directly after every TG. By contrast, if the trustee did *not* establish information sharing, each trustor knows only the outcomes of her own past TGs. I will refer to the continuation of Γ after the trustee did *not* establish information sharing in

period 0.2 as (*continuation game*) Γ^{-} and to the continuation of the game after the trustee established information sharing as (*continuation game*) Γ^{+}. Which continuation game is played, that is, whether the trustee has established information sharing, is common knowledge.

Payoffs

The payoffs in any TG are as follows. If the trustor i who is at play does not place trust, trustor i and the trustee earn the payoffs P_1 and P_2, respectively. If i places trust, i gets $R_1 > P_1$ if the trustee honors trust and $S_1 < P_1$ if the trustee abuses trust. The trustee gets $R_2 > P_2$ if he honors trust and $T_2 > R_2$ if he abuses trust. These are the material payoffs and I assume that trustors and trustees of the opportunistic type are exclusively concerned with own material payoffs. An opportunistic trustee would thus maximize his utility in the *focal* TG by abusing trust. Friendly trustees, however, have non-material motivations such that if i places trust, a friendly trustee maximizes his utility by honoring trust. I consider two scenarios with respect to the non-monetary motives of friendly trustees.

- In Γ^{ga}, a trustee of the friendly type is a guilt-avoiding trustee. A guilt-avoiding trustee has no incentive to abuse i's trust because if he abuses trust, he suffers from an internal sanction θ that offsets the material incentive to abuse trust ($T_2 - \theta < R_2$).
- In Γ^{rs}, a trustee of the friendly type is a reward-seeking trustee. A reward-seeking trustee has no incentive to abuse i's trust because he receives an internal reward υ if he honors trust such that $T_2 < R_2 + \upsilon$.

An actor's total payoff for Γ (i.e. Γ^{ga} as well as Γ^{rs}) is the sum of the undiscounted payoffs that the actor earns in the TGs in the periods 1 to mN minus—in case of the trustee—the cost c of an investment in establishing information sharing if such an investment has been pledged.

Analysis of the game

The aim of the analysis is to identify under what conditions there exist equilibria in which the trustee invests in establishing information sharing. A rational trustee will weigh the cost of investment c against the expected ''return on investment,'' namely, his expected payoff in Γ^{+} minus his expected payoff in Γ^{-}. Section ''Behavior and payoffs in the TGs'' specifies the equilibria and expected payoffs of Γ^{+} and Γ^{-}. This section shows

that network effects can lead to more trust and higher earnings in Γ^+ than in Γ^-. Furthermore, we shall see that the trustee's expected payoff also tends to be higher if, at the beginning of the TGs, the trustors are more convinced that the trustee is of the friendly type. Section ''Investments in establishing information sharing'' investigates under what conditions a trustee invests in establishing information sharing to benefit from the network effects or to convince the trustors that he is of the friendly type.

Throughout, I focus on *sequential* equilibria—a refinement of Nash equilibrium for games with incomplete information (Kreps and Wilson, 1982b; see Rasmusen, 1994: Chap. 6, for a textbook). For the analysis of investments in information sharing, I additionally employ the *passive conjectures* refinement (Rasmusen, 1994; Rubinstein, 1985), which will be explained below. I will often refer to sequential equilibria and sequential equilibria involving the passive conjecture for brevity as ''equilibria.''

Behavior and payoffs in the TGs

We can lend from earlier work for the analysis of Γ^- and Γ^+. Camerer and Weigelt (1988) have analyzed reputation building in sequential equilibria in finitely repeated TGs with incomplete information (see also Anderhub et al., 2002; Bower et al., 1997). In a nutshell, if one trustor and one trustee play a finitely repeated TG together, there is typically a sequential equilibrium such that trust is placed and honored in early periods and, then, there is an ''endgame'' in which trust may break down. In early periods also a trustee of the opportunistic type typically honors trust, balancing the short-term incentive to abuse trust and the long-term benefit of having a reputation for being of the friendly type. Toward the end of the game, however, the long-term benefit of having a good reputation decreases, the opportunistic trustee begins to abuse trust with positive probability, and trust may break down.

If there is no information sharing (Γ^-), this equilibrium applies to the N TGs of each trustor separately, as if there were no other trustors who interact with the trustee (cf. Buskens, 2003; Frey et al., 2015). By contrast, in Γ^+, where the trustors can learn from each other's experiences with the trustee, this sequential equilibrium applies to the total number of mN TGs as if it was one single trustor who plays *all* these TGs (Buskens, 2003; Camerer and Weigelt, 1988; Frey et al., 2015).[4] Lemma 1 in Appendix 1 specifies the sequential equilibria of the continuation games Γ^- and Γ^+ formally. In the following, I describe the sequential equilibrium in more detail, focusing on the interactions between one trustor and the trustee in

Γ^-. I also describe how the equilibrium applies to Γ^+ and specify the expected payoffs of the trustee for Γ^- and Γ^+. Note that the sequential equilibria of Γ^- and Γ^+ are the same in Γ^{ga} and Γ^{rs}. The *reason* for which a friendly trustee always honors trust is irrelevant for the equilibrium of the TGs.[5]

We will need some further notation. Let π^- and π^+ denote the beliefs of the trustors about the trustee's type that enter the continuation games Γ^- and Γ^+, respectively. In addition, let $RISK := (P_1 - S_1)/(R_1 - S_1)$ measure the risk a trustor incurs when placing trust and let $TEMP := (T_2 - R_2)/(T_2 - P_2)$ measure an opportunistic trustee's temptation to abuse trust (see Snijders, 1996).

Behavior and payoffs in Γ^-. If there is no information sharing, what happens in the TGs between one trustor and the trustee is independent of what happens in the other trustors' TGs. In Γ^-, the sequential equilibrium in each trustor i's TGs is, therefore, the same as if there was only one trustor playing N TGs with the trustee. Anderhub et al. (2002) and Camerer and Weigelt (1988) provide detailed and accessible descriptions of this equilibrium. The equilibrium evolves typically over three phases. First, trustor i places trust and the trustee honors trust until trustor i and the trustee have τ^- TGs left to play together, where τ^- is the smallest integer for which it holds that $\pi^- \geq RISK^{\tau^-}$, that is

$$\tau^- := \left\lceil \frac{\log \pi^-}{\log RISK} \right\rceil \text{ for } 0 < \pi^- < 1 \tag{1}$$

Then, in the next TG of trustor i, the trustee—if he is of the opportunistic type—starts to randomize. Still one TG later, trustor i begins to randomize too, placing trust with probability *TEMP*. In this second phase, the "randomization phase," trustor i and the trustee (if he is of the opportunistic type) randomize until i does not place trust or until the trustee abuses i's trust.[6] After the first instance that i did not place trust or that the trustee abused i's trust, the third and final phase starts: i does not place trust anymore. If i still places trust in her last TG, an opportunistic trustee abuses trust.

The equilibrium does not always evolve over all these phases. If the trust problem is very severe, there may be no trust at all. As equation (1) shows, τ^- tends to be larger (the randomization tends to start earlier) if π^- is smaller or *RISK* is larger. If π^- is so small or *RISK* so large that $\tau^- = N$, the opportunistic trustee randomizes already in the first TG with trustor i. If π^- and *RISK* are even such that $\tau^- > N$, i never places trust (because,

given the parameters, there are not enough interactions with trustor i for the trustee to start building a reputation).

The equilibrium also does not exhibit a randomization phase if $\pi^- = 0$ or $\pi^- = 1$. In these cases, it is as if trustor i had complete information about the trustee's incentives, and τ^- is not specified by equation (1). If i is totally convinced that the trustee is of the opportunistic type ($\pi^- = 0$), the logic of backward induction implies that i should never place trust, irrespectively of how long the game is. Therefore, we say that $\tau^- = \infty$ if $\pi^- = 0$. In contrast, if $\pi^- = 1$, i should place trust even in a one-shot TG. In the repeated game, i will then in equilibrium place trust in every TG. If the trustee is indeed of the friendly type, he always honors trust. If—contrary to i's belief—the trustee is of the opportunistic type, he waits with abusing trust until i's last TG. This is the same equilibrium course of play as if $RISK < \pi^- < 1$, and thus, by equation (1), $\tau^- = 1$. We, therefore, say that $\tau^- = 1$ if $\pi^- = 1$.

We can now specify the expected payoffs of friendly and opportunistic trustees in Γ^{ga-} and Γ^{rs-}.

Proposition 1. The expected payoffs of a friendly trustee in Γ^{ga-} $(U_F^{\Gamma^{ga-}})$ *and* Γ^{rs-} $(U_F^{\Gamma^{rs-}})$ *are*

$$U_F^{\Gamma^{ga-}} = \begin{cases} m\big((N-\tau^-)R_2 + \tau^- P_2 \\ \quad + (T_2 - P_2)(1 - TEMP^{\tau^-})\big) & \text{if } \tau^- \leq N \\ mNP_2 & \text{if } \tau^- > N \end{cases}$$

$$U_F^{\Gamma^{rs-}} = \begin{cases} m\big((N-\tau^-)(R_2 + \upsilon) + \tau^- P_2 \\ \quad + (T_2 - P_2)\frac{R_2 + \upsilon - P_2}{R_2 - P_2}(1 - TEMP^{\tau^-})\big) & \text{if } \tau^- \leq N \\ mNP_2 & \text{if } \tau^- > N \end{cases}$$

The expected payoffs of an opportunistic trustee in Γ^{ga-} $(U_O^{\Gamma^{ga-}})$ *and* Γ^{rs-} $(U_O^{\Gamma^{rs-}})$ *are*

$$U_O^{\Gamma^{ga-}} = U_O^{\Gamma^{rs-}} = \begin{cases} m\big((N-\tau^-)R_2 + T_2 + (\tau^- - 1)P_2\big) & \text{if } \tau^- \leq N \\ mNP_2 & \text{if } \tau^- > N \end{cases}$$

The proofs of all propositions are in Appendix 1. For the scenario that $\tau^- \leq N$, one can think of the trustee's expected payoff for the TGs with each of the m trustors as consisting of two components. The first component pertains to the phase during which trust is placed and honored with certainty. In Γ^{ga}, this is $(N - \tau^-)R_2$ for both trustees; in Γ^{rs}, this is likewise $(N - \tau^-)R_2$ for the opportunistic trustee and it is $(N - \tau^-)(R_2 + \upsilon)$ for the friendly trustee. The second component, the reminder of the

formulas in Proposition 1, pertains to the τ^- last TGs that the trustee plays with each trustor i, the second and third phase of the equilibrium (see the proof of Proposition 1 for details).

Behavior and payoffs in Γ^+. In Γ^+, network effects facilitate trust and trustworthiness. The trustors can learn from each other's experiences. Therefore, the trustee has to take into account that his choice in one TG will affect the future choices of *all* trustors. Given the assumption that information is always shared truthfully, the trustors should condition their behavior on third-party information in the same manner as on personal experience (cf. Bolton et al., 2004). An opportunistic trustee's incentive to honor trustor i's trust in a given period n is, therefore, the same as if he played *all* remaining $mN - n$ TGs with that trustor (even if he plays some or all of these TGs with different trustors). Hence, the equilibrium that applies to the interactions between *one* trustor i and the trustee in Γ^- applies to the interactions between *all* trustors and the trustee in Γ^+ (Buskens, 2003; Camerer and Weigelt, 1988; Frey et al., 2015).

Specifically, in Γ^+, trust is placed and honored until there are *in total* τ^+ TGs left, where τ^+ is calculated as τ^- using equation (1), but with π^+ instead of π^- (hence, if $\pi^+ = \pi^-$, then $\tau^+ = \tau^-$). Next, the randomization begins. And after the first instance that the trustee abuses the trust of *one* trustor or that *one* trustor does not place trust, *no* trustor ever places trust again. Analogous to the situation in Γ^-, the trustors never place trust in Γ^+ if $\tau^+ > mN$, and we also assume that $\tau^+ = \infty$ if $\pi^+ = 0$ and $\tau^+ = 1$ if $\pi^+ = 1$. Proposition 2 specifies the expected payoffs of a trustee in Γ^{ga+} and Γ^{rs+}.

Proposition 2. The expected payoffs of a friendly trustee in Γ^{ga+} $(U_F^{\Gamma^{ga+}})$ *and* Γ^{rs+} $(U_F^{\Gamma^{rs+}})$ *are*

$$U_F^{\Gamma^{ga+}} = \begin{cases} (mN - \tau^+)R_2 + \tau^+ P_2 + (T_2 - P_2)(1 - TEMP^{\tau^+}) & \text{if } \tau^+ \leq mN \\ mNP_2 & \text{if } \tau^+ > mN \end{cases}$$

$$U_F^{\Gamma^{rs+}} = \begin{cases} (mN - \tau^+)(R_2 + \upsilon) + \tau^+ P_2 & \text{if } \tau^+ \leq mN \\ \quad + (T_2 - P_2)\frac{R_2 + \upsilon - P_2}{R_2 - P_2}(1 - TEMP^{\tau^+}) & \\ mNP_2 & \text{if } \tau^+ > mN \end{cases}$$

The expected payoffs of an opportunistic trustee in Γ^{ga+} $(U_O^{\Gamma^{ga+}})$ *and* Γ^{rs+} $(U_O^{\Gamma^{rs+}})$ *are*

$$U_O^{\Gamma^{ga+}} = U_O^{\Gamma^{rs+}} = \begin{cases} (mN - \tau^+)R_2 + T_2 + (\tau^+ - 1)P_2 & \text{if } \tau^+ \leq mN \\ mNP_2 & \text{if } \tau^+ > mN \end{cases}$$

The comparison of equilibrium behavior and expected payoffs in Γ^- and Γ^+ reveals the network effects. All else equal (including the trustors' belief about the trustee's type at the beginning of the TGs), trust tends to be placed and honored during more TGs and the trustee earns more in Γ^+ than in Γ^-. If trust is also possible in Γ^-, the trustee can avoid having the randomization phase with each trustor separately in Γ^+. If trust is not possible in Γ^-, the network effects can make a phase with honored trust possible in Γ^+. Another important observation is that the trustee will also be trusted more and earn a higher payoff if the trustors are at the beginning of the TGs more convinced that they are dealing with a friendly trustee (because a higher belief π^- and π^+ implies a smaller τ^- and τ^+, respectively).

Investments in establishing information sharing

In this section, I establish under what conditions there exist equilibria in which the trustee establishes information sharing between the trustors. Let ρ_F and ρ_O denote the probability with which friendly and opportunistic trustees establish information sharing, respectively. There are only two combinations of ρ_F and ρ_O that can be part of an equilibrium that involves an investment of the trustee. These are (1) both trustees invest with probability 1 (i.e. $\rho_F = \rho_O = 1$) and (2) the friendly trustee invests while the opportunistic trustee does not invest ($\rho_F = 1$ and $\rho_O = 0$).

For the analysis of investments in information sharing, it is important to realize that a trustee's expected payoffs for Γ^- and Γ^+ depend on the trustors' beliefs entering these continuation games and that these beliefs depend on the investment probabilities ρ_F and ρ_O. Bayes' rule implies that $\pi^- = \big((1-\rho_F)\pi\big)/\big((1-\rho_F)\pi + (1-\rho_O)(1-\pi)\big)$ and $\pi^+ = \rho_F\pi/\big(\rho_F\pi + \rho_O(1-\pi)\big)$. Hence, to assess whether some combination of ρ_F and ρ_O can be part of a sequential equilibrium, one needs to check whether the trustees maximize their expected payoffs by investing with the assumed probabilities ρ_F and ρ_O, *given* the beliefs π^- and π^+ that are implied by these probabilities.[7]

In the following, I establish under what conditions it is part of an equilibrium that both trustees establish information sharing ($\rho_F = \rho_O = 1$; scenario (1)) or that only the friendly trustee establishes information sharing ($\rho_F = 1$ and $\rho_O = 0$; scenario (2)). We shall see that the trustee's investment promotes trust for different reasons in these two scenarios. In brief, in ''pooling equilibria'' in which $\rho_F = \rho_O = 1$, the investment enables network effects but does not affect the beliefs of the trustors. By contrast, in ''separating equilibria'' in which only the friendly trustee invests, the investment

leads to higher payoffs for the trustee in Γ^+ exclusively because it signals to the trustors' that they are dealing with a friendly trustee.

Investments in information sharing to allow network effects. In equilibria in which the trustee invests in establishing information sharing regardless of his type ($\rho_F = \rho_O = 1$), the investment enables the network effects that promote trust and trustworthiness, but it does not signal intrinsic trustworthiness. If $\rho_F = \rho_O = 1$, Bayes' rule implies that π^+ equals the prior probability π. If, unexpectedly, the trustee deviates from a conjectured equilibrium in which $\rho_F = \rho_O = 1$ and does not invest, the trustors do likewise not learn anything because neither trustee should ever do this. π^- is not determined by Bayes' rule and, in principle, the trustors could hold any ''out-of-equilibrium'' belief $0 \leq \pi^- \leq 1$. However, I employ the passive conjecture refinement to avoid constructing equilibria that require beliefs π^- that are difficult to justify.[8] That is, I assume that the trustors do not change their beliefs upon observing a deviation from a conjectured equilibrium in period 0.2.[9] Thus, if $\rho_F = \rho_O = 1$, $\pi^- = \pi^+ = \pi$ and, hence, $\tau^- = \tau^+$. The investment does not make the trustors more or less optimistic about the trustee's type and it can benefit the trustee only due to the network effects. To simplify the discussion of the conditions for the existence of equilibria in which $\rho_F = \rho_O = 1$, I use τ without a superscript $-$ or $+$ to denote simultaneously τ^- and τ^+.

Γ has an equilibrium in which $\rho_F = \rho_O = 1$ if c does not exceed the benefit that the network effects yield for the trustee who benefits least from these effects. That is, there exists an equilibrium in which $\rho_F = \rho_O = 1$ if $c \leq \bar{c} := \min(U_F^{\Gamma^+(\tau)} - U_F^{\Gamma^-(\tau)}, U_O^{\Gamma^+(\tau)} - U_O^{\Gamma^-(\tau)})$, where, for example, $U_F^{\Gamma^+(\tau)}$ denotes the expected payoff of a friendly trustee for Γ^+ given the τ (i.e. τ^+) that results from $\pi^+ = \pi$.

Proposition 3. Γ^{ga} *has a sequential equilibrium in which* $\rho_F = \rho_O = 1$ *if and only if* $c \leq \bar{c}^{ga}$, *where*

$$\bar{c}^{ga} = \begin{cases} (m-1)\big(\tau(R_2 - P_2) - (T_2 - P_2)\big) & \text{if } \tau \leq N \\ (mN - \tau)(R_2 - P_2) + (T_2 - P_2)(1 - TEMP^{\tau}) & \text{if } N < \tau \leq mN \end{cases}$$

Γ^{rs} *has a sequential equilibrium in which* $\rho_F = \rho_o = 1$ *if and only if* $c \leq \bar{c}^{rs}$, *where*

$$\bar{c}^{rs} = \begin{cases} (m-1)\big(\tau(R_2 - P_2) - (T_2 - P_2)\big) & \text{if } \tau \leq N \\ (mN - \tau)(R_2 - P_2) + (T_2 - P_2) & \text{if } N < \tau \leq mN \end{cases}$$

Proposition 3 shows that $\bar{c}$ is similar for Γ^{ga} and Γ^{rs} and distinguishes two scenarios: $\tau \leq N$ and $N < \tau \leq mN$. If $\tau \leq N$, the trustors place trust in some TGs also in Γ^{-}. However, establishing information sharing can payoff for the trustee because the network effects enable the trustee to have the randomization phase just once rather than with each of the m trustors separately. Therefore, a trustee earns approximately $(m-1)\tau(R_2 - P_2)$ more in Γ^{+} than in Γ^{-}. An opportunistic trustee benefits somewhat less from the network effects than a friendly trustee because he has the opportunity to abuse trust m times in Γ^{-} but only once in Γ^{+}. Hence, if $\tau \leq N$, $\bar{c}$ equals how much the opportunistic trustee benefits from the network effects, which is the same in Γ^{ga} and Γ^{rs}.

If $N < \tau \leq mN$, the trustors never place trust if the trustee does not establish information sharing (because, given the parameters, the sanctioning potential of one trustor alone is not sufficient to motivate an opportunistic trustee to start building a reputation). However, the network effects make honored trust possible in Γ^{+}. In this case, a trustee's return on investment equals approximately his benefit from trust being placed and honored with certainty in $mN - \tau$ TGs in Γ^{+} compared to never being trusted in Γ^{-}, that is, approximately $(mN - \tau)(R_2 - P_2)$. For the friendly trustee, the return on investment is smaller in Γ^{ga} than in Γ^{rs}. Therefore, $\bar{c}$ is somewhat smaller in Γ^{ga} than in Γ^{rs} ($\bar{c}$ is by $(T_2 - P_2)TEMP^{\tau}$ smaller in Γ^{ga} (where $\bar{c}$ equals the friendly trustee's return on investment) than in Γ^{rs} (where $\bar{c}$ equals the opportunistic trustee's return on investment)).

The next proposition establishes how parameter changes affect $\bar{c}$. All parameters affect $\bar{c}$ in the same manner in Γ^{ga} and Γ^{rs}, apart from a minor difference in the effect of changes in R_2. I continue using $\bar{c}$ without a superscript to refer to $\bar{c}^{ga}$ and $\bar{c}^{rs}$ simultaneously.

Proposition 4. The maximum cost $\bar{c}$ for which Γ has a sequential equilibrium in which $\rho_O = \rho_F = 1$ depends on the parameters of the game as follows:

(1)
- *If $\tau \leq N$, $\bar{c}$ increases weakly if RISK increases (i.e. if P_1 increases or S_1 or R_1 decreases) or if π decreases.*
- *If $N < \tau \leq mN$, $\bar{c}$ decreases weakly if RISK increases (i.e. if P_1 increases or S_1 or R_1 decreases) or if π decreases.*

(2)
- *If $\tau \leq mN$, $\bar{c}^{ga}$ increases in R_2.*
- *If $\tau \leq mN$, $\bar{c}^{rs}$ increases in R_2.*
- *If $1 < \tau \leq mN$, $\bar{c}$ decreases in P_2.*
- *If $\tau \leq N$, $\bar{c}$ decreases in T_2.*

(3)
- *If* $N < \tau \leq mN$, $\bar{c}$ *increases in* T_2.
- *If* $N < \tau \leq mN$, $\bar{c}$ *increases in* m.
- *If* $N < \tau \leq mN$, $\bar{c}$ *increases in* N.

List (1) of Proposition 4 concerns the effects of changes in π and $RISK = (P_1 - S_1)/(R_1 - S_1)$, that is, the probability of a friendly trustee and the risk a trustor incurs when placing trust. It shows that a trustee's incentive to establish information sharing changes in an inverse U-shape over π and *RISK*. The trustee's incentives to invest are small for small trust problems (large π or small *RISK*) where there is a lot of trust even without information sharing and, hence, information sharing increases trust only marginally. The incentives to invest are also small if trust problems are very large (small π or large *RISK*) and honored trust is hardly attainable even with information sharing. However, the incentives to invest are large for trust problems of an intermediate size in which the network effects make a considerable difference for the behavior of trustors and trustee.

To understand this inverse U-shape in more detail, recall that if π increases or *RISK* decreases, the randomization phase tends to start earlier (τ tends to decrease; see equation (1)). Now assume that π is not too small and *RISK* is not too large such that trust is also possible without information sharing ($\tau \leq N$). In that case, establishing information sharing enables the trustee to have the randomization phase just once rather than with each trustor separately. Obviously, this is more valuable if the randomization phase is longer due to a smaller π or larger *RISK*. More formally, if $\tau \leq N$, the network effects enable the trustee to benefit from trust being placed and honored with certainty in $(m-1)\tau$ additional TGs, which is more valuable if τ is larger. Hence, as long as π and *RISK* are such that $\tau \leq N$, $\bar{c}$ increases if π or *RISK* changes such that τ increases. The effects are opposite once π and *RISK* are such that trust is no longer possible without information sharing. If $N < \tau \leq mN$, $\bar{c}$ becomes smaller if π or *RISK* further changes such that τ increases. An increase in τ then leads to less trust in Γ^+ while there was already no trust in Γ^- before the increase in τ.

List (2) of Proposition 4 concerns the TG payoffs of the trustee. $\bar{c}$ is larger if the benefit a trustee derives from honored trust compared to no trust $(R_2 - P_2)$ is larger. Furthermore, $\bar{c}$ decreases in the payoff an opportunistic trustee earns when abusing trust (T_2) if $\tau \leq N$, but increases in T_2 if $N < \tau \leq mN$.[10] Note, furthermore, that if $\tau \leq N$, there may be no equilibrium in which $\rho_F = \rho_O = 1$ even for an arbitrarily small cost $c > 0$. This occurs if the loss an opportunistic trustee incurs from having only once the

opportunity to abuse trust in Γ^+ rather than m times in Γ^- is large while his benefit from an extended phase of trust and trustworthiness is small (if $\tau \leq N$ and $\tau(R_2 - P_2) < T_2 - P_2$, $\bar{c}$ is negative; see Proposition 3).

List (3) of Proposition 4 shows that $\bar{c}$ increases in the number m of trustors and the number N of TGs played with each trustor if $N < \tau \leq mN$. If $N < \tau \leq mN$, an increase in m implies that the trustee benefits from honored trust in N additional TGs in Γ^+ while he will not be trusted in these N additional TGs in Γ^-. If $\tau < N$, $\bar{c}$ may also increase in m because if there is one additional trustor, trust is placed and honored with certainty in N additional TGs in Γ^+ and only in $N - \tau$ additional TGs in Γ^-. However, if $\tau < N$ and $\bar{c}$ is negative because $\tau(R_2 - P_2) < T_2 - P_2$, $\bar{c}$ decreases even further if m increases because the opportunistic trustee benefits more from having the chance to abuse the additional trustor's trust than from an extended phase of honored trust.

An increase in N also leads to an increase in $\bar{c}$ if $N < \tau \leq mN$. If $N < \tau \leq mN$, an increase in N means adding one TG for each trustor in which the trustee is not trusted in Γ^- but in which he benefits from honored trust in Γ^+. However, if trust is also possible in Γ^- (i.e. $\tau < N$), $\bar{c}$ does not change in N. An increase in N then means adding one TG for each trustor in which the trustee benefits from honored trust in Γ^- as well as in Γ^+.

Summarizing, the condition for the existence of pooling equilibria in which the trustee establishes information sharing to benefit from network effects is almost the same in Γ^{ga} and Γ^{rs} (i.e. if friendly trustees are guilt-avoiding trustees and if they are reward-seeking trustees). Important results are that the maximum cost $\bar{c}$ for which there are such equilibria is high especially if (1) the size of the trust problem is intermediate—π and *RISK* are neither too small nor too large—such that network effects bring about a considerable increase in honored trust and if (2) the trustee benefits a lot from honored trust compared to no trust ($R_2 - P_2$ is large).

Investments in information sharing as signals of intrinsic trustworthiness. In equilibria in which only the friendly trustee invests in establishing information sharing ($\rho_F = 1$ and $\rho_O = 0$), the trustee's investment is a perfectly discriminating signal of intrinsic trustworthiness. After observing the trustee's investment decision, the trustors know whom they are dealing with. They always place trust if the trustee invested and never place trust if the trustee did not invest. Formally, $\rho_F = 1$ and $\rho_O = 0$ together imply $\pi^- = 0$ and $\pi^+ = 1$ and, hence, $\tau^- = \infty$ and $\tau^+ = 1$. The results concerning such separating equilibria are quite different for Γ^{ga} and Γ^{rs} and, therefore, presented separately. For the game with guilt-avoiding trustees, we have the following proposition:

Proposition 5. *In Γ^{ga}, there cannot exist a sequential equilibrium in which $\rho_F = 1$ and $\rho_O = 0$.*

An equilibrium in which $\rho_F = 1$ and $\rho_O = 0$ requires that the benefit from always being trusted compared to never being trusted is larger for a friendly trustee than for an opportunistic trustee. However, this is not the case in Γ^{ga}. Both trustees earn mNP_2 if the trustors never place trust. If the trustors always place trust, the opportunistic trustee earns more than the friendly trustee, namely, $(mN - 1)R_2 + T_2 > mNR_2$ (because he can abuse trust in the last TG without feeling remorse). Hence, Γ^{ga} cannot have an equilibrium in which $\rho_F = 1$ and $\rho_O = 0$.

A remark is in order. Γ^{ga} could have an equilibrium in which $\rho_F = 1$ and $\rho_O = 0$ if friendly trustees could establish information sharing at lower costs than opportunistic trustees.[11] Assuming that different types face different costs of sending some signal is not uncommon (e.g. Aksoy and Gambetta, 2016; Gintis et al., 2001; Patel, 2012). For settings with relatively few trustors, one could defend such an assumption based on arguments along the line that a trustee with social preferences (a friendly trustee) is more adept at bringing people together than a trustee who is exclusively interested in material outcomes or that a friendly trustee even derives some pleasure from doing this. However, I here focus on whether a trustee can signal his intrinsic trustworthiness by establishing information sharing when trustee types differ exclusively in their payoffs in the TG. My last proposition establishes that this is possible in Γ^{rs} where a friendly trustee receives an internal reward υ every time he honors trust.

Proposition 6. *In Γ^{rs}, there exists a sequential equilibrium in which $\rho_F = 1$ and $\rho_O = 0$ (and in which the trustee is always trusted after having established information sharing but never trusted otherwise) if and only if $mN(R_2 - P_2) + T_2 - R_2 \leq c \leq mN(R_2 + \upsilon - P_2)$.*

In Γ^{rs}, a friendly trustee benefits always more from being trusted in some TG than an opportunistic trustee, irrespectively of whether the latter honors or abuses trust (because $R_2 + \upsilon > T_2$). It can, therefore, be worthwhile for a friendly trustee to invest in information sharing to signal his type and induce the trustors to place trust while it does not payoff for an opportunistic trustee to mimic the friendly trustee by investing, too. To deter mimicry by the opportunistic trustee, the cost c must be ''rather high.'' The comparison of Propositions 3 and 6 shows that the lower bound on c for the existence of a separating equilibrium is higher than the maximum cost for which Γ^{rs} has a pooling equilibrium in which $\rho_F = \rho_O = 1$

and in which the investment affects the trustors' behavior less strongly.[12] Proposition 6 shows that the lower bound for c for which there exists an equilibrium in which $\rho_F = 1$ and $\rho_O = 0$ increases in m, N, $R_2 - P_2$, and T_2. On the other hand, c must also be small enough such that it pays off for a friendly trustee to signal his type by investing. The maximum cost c for which this is the case increases in m, N, $R_2 - P_2$, and υ, as can be inferred from Proposition 6. The cost range for which Γ^{rs} has a separating equilibrium is larger if m or N is larger (because a friendly trustee can then receive the internal reward υ more often), if υ is larger, and if an opportunistic trustee's benefit from abusing trust ($T_2 - R_2$) is smaller.

Summarizing, a trustee's investment in establishing information sharing can signal that he is of the friendly type if friendly trustees are trustworthy because they derive pleasure from good behavior (as in Γ^{rs}) but not if they are trustworthy because they would suffer an internal sanction if abusing trust (as in Γ^{ga}). In the former case, signaling equilibria can exist especially if the cost of investment c is high, the trustee interacts many times with many trustors, friendly trustees derive a large internal reward from honoring trust, and opportunistic trustees cannot gain too much from abusing trust.

Conclusion and discussion

In this article, a game-theoretic model has been developed that advances our understanding of the emergence and effects of institutions for information sharing in trust situations. The model allows for a simultaneous investigation of investments in setting up an institution for information sharing and the effects of such an institution. It focuses on investments in information sharing by a trustee, and it captures two mechanisms by which a trustee's investment can promote trust. First, the trustee's investment can allow network effects that facilitate trust and trustworthiness. Second, the trustee's investment can serve as a credible signal of intrinsic trustworthiness.

A core result is the prediction that there is an inverse U-shape in the relation between the size of the trust problem and the incentives for establishing information sharing in order to allow network effects. A trustee should be particularly likely to establish information sharing to enable network effects if the size of the trust problem is intermediate (if there is an intermediate probability π of interacting with an intrinsically trustworthy trustee and if the risk that a trustor incurs when placing trust (*RISK*) is neither very small nor very large). This could suggest, for example, that someone who sells an ordinary consumer good in a brick and mortar market will *not* invest in setting up an institution for information sharing because

the trust problem is too small. However, someone who sells the same good over the Internet may invest in a reputation system because buyers are more afraid of fraud in online trade.

This inverse U-shape prediction is in line with the results of Frey et al. (2015) and Raub et al. (2013). Frey et al. (2015) report qualitatively the same effects of changes in π and *RISK* for the scenario that the *trustors*, rather than the trustee, can establish information sharing. In the Raub et al. (2013) model—a game with indefinite repetition and complete information on all actors' incentives—network effects depend exclusively on the incentives of trustees (and not at all on the trustors' incentives). Still, this alternative model, which also covers social dilemmas other than the TG, leads to a similar conclusion: investments in information sharing should be more likely if the incentives to take advantage of others are neither very small nor very large. This article thus reinforces the testable prediction of an inverse U-shape in the relation between the size of the trust problem and the incentives to establish information sharing in order to allow network effects.

A second key result of this study is that a trustee's investment in information sharing can serve as a credible signal of intrinsic trustworthiness. The analysis suggests that a trustee is particularly likely to establish information sharing to signal his intrinsic trustworthiness if he derives a large internal reward from good behavior, if there are many trustors and many interactions with each trustor, and if the investment cost is high. The analysis, furthermore, shows that signaling can only occur if intrinsically trustworthy trustees derive a larger benefit from being trusted than trustees that have a short-term incentive to abuse trust.

The article also offers new predictions on the *effects* of institutions for information sharing. Earlier work focused exclusively on network effects. The result that a trustee's investment in information sharing can serve as a signal of intrinsic trustworthiness suggests that information sharing can affect behavior more strongly than previous work predicts. Network effects tend to affect behavior gradually, as in my model in which they lead to honored trust in some more TGs. Signaling can have more swiping effects. In separating equilibria in which only the intrinsically trustworthy trustee invests, the trustors never place trust if the trustee does not invest while they place trust throughout if the trustee does invest. This suggests that information sharing can promote trust more strongly if it was established by the trustee than if it is given exogenously. Another implication is that a trustee's investment in information sharing may have a particularly strong effect on trust if the costs of investment are high.

The results on the conditions for separating equilibria also generalize beyond the context of investments in information sharing. In a separating equilibrium, the trustee's investment is a ''wasteful signal.'' After having observed the investment, the trustors know for sure that they are dealing with an intrinsically trustworthy trustee. There is no more need for the network effects. Therefore, the trustee could signal his type by ''burning resources'' in any other way than by creating a context that *would* promote trust *if* the trustors were uncertain about the trustee's type. In this sense, the analysis establishes under what conditions a trustee who interacts N times with m trustors sends *any* costly signal to convey his intrinsic trustworthiness. The results on how parameters affect the restrictiveness of the conditions for the existence of such separating equilibria are in line with results of Przepiorka and Diekmann (2013) for infinitely repeated TGs.

Although a trustee could take various actions to signal his intrinsic trustworthiness, I believe that establishing information sharing is a particularly attractive option for three reasons. First, trustors have an interest in acquiring information about a trustee's reputation. It is therefore, for example, unlikely that it escapes the attention of potential buyers that an online seller invested in a reputation system. At least, it is probably more likely that they overlook a note that reports a charitable donation (see Fehrler and Przepiorka, 2013, 2016; Milinski et al., 2002 for studies on charitable giving as a signaling device). That is, an investment in establishing information sharing might be a signal with high ''broadcast efficiency'' (Bliege Bird and Smith, 2005; Gintis et al., 2001). Second, there may be a chance that trustors do not understand a trustee's action as a signal of intrinsic trustworthiness. If the signal fails to convey its intended message, the trustee can still benefit from network effects if he tried to signal his type by establishing information sharing. Finally, trustors might not be totally convinced of the trustee's intrinsic trustworthiness also after having observed the signal (for example, due to heterogeneity in the cost of sending the signal that can make it possible for an opportunistic trustee to afford sending the signal). Network effects can then further increase the level of trust after signaling. These considerations highlight the importance of studying how signaling can drive investments in information sharing. However, these considerations also suggest that it may be a limitation that in the presented model an investment in information sharing facilitates exchange *either* because it enables network effects *or* because it signals intrinsic trustworthiness. Future work could modify the presented model to investigate how signaling and network effects can complement one another.

Acknowledgements

The author thanks Vincent Buskens and Werner Raub for comments and suggestions at various stages of the research process and Jacob Dijkstra and Victor Stoica for comments on an earlier version of the manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Netherlands Organization for Scientific Research (NWO, Graduate Program Grant 2008/2009 for the ICS).

Notes

1. To facilitate identifying the actors, I use female pronouns for trustors and male pronouns for trustees.
2. See, for example, Camerer (2003: Chap. 2), Cooper and Kagel (2013: Section III.A), Fehr and Fischbacher (2003) and Johnson and Mislin (2011) for overviews on trust and trustworthiness in experimental studies.
3. I assume intrinsic trustworthiness to be a stable trait rather than a state. I assume that a trustee is either of the friendly type or of the opportunistic type and I do not take into account that non-material motives may depend on situational factors that frame the specific decision situation (e.g. Andreoni, 1995; Lindenberg and Steg, 2007) or on earlier choices of other players (e.g. Attanasi et al., 2015; Battigalli and Dufwenberg, 2007). Note, furthermore, that the model is, in principle, not restricted to the scenario that friendly trustees are trustworthy due to non-material motivations. A friendly trustee could also be trustworthy because he simply does not have an attractive option to abuse trust.
4. Already the initial contribution of Camerer and Weigelt (1988) shows the theoretical equivalence of the situation in which a trustee plays repeatedly with one trustor and the situation that there are different trustors who share information. In the experiment of Camerer and Weigelt, one trustee played with a new trustor every period, but the trustors always knew the entire history (as in Γ^{+}). The sequential equilibrium predictions that Camerer and Weigelt tested were the same as if the trustee played with the same trustor throughout.
5. In the seminal works on reputation building in sequential equilibria, the player type who never takes advantage of his opponent (the ''friendly player'') was referred to as ''irrational player'' and the motivation for his behavior was left unspecified (see Kreps et al., 1982).
6. In the randomization phase, trustor i becomes more confident that the trustee is of the friendly type with every additional time that trust is honored, because a friendly trustee always honors trust while an opportunistic trustee abuses trust with positive probability (the trustor learns in the sense of Bayesian

updating). On the other hand, the risk of trust abuse increases over the periods of the randomization phase because the probability with which an opportunistic trustee abuses trust increases toward the end of the game. These two effects cancel each other out such that a trustor is in every period indifferent between placing and not placing trust.

7. There are, in principle, two further combinations of ρ_F and ρ_O that give a positive probability of an investment in information sharing. However, these cannot be part of an equilibrium. First, if only the opportunistic trustee invests ($\rho_F = 0$ and $\rho_O = 1$), the opportunistic trustee would thereby reveal himself as being of the opportunistic type and, consequently, never be trusted. Formally, $\rho_F = 0$ and $\rho_O = 1$ together imply $\pi^- = 1$ and $\pi^+ = 0$ and, hence, $\tau^- = 1$ and $\tau^+ = \infty$. However, if $\tau^- = 1$ and $\tau^+ = \infty$, the opportunistic trustee would better not invest. Second, it could be that one or both trustees randomize between investing and not investing. This would require that c equals a trustee's return on investment precisely. However, this can hold only for very specific parameter constellations because the trustees' expected payoffs for Γ^- and Γ^+ do not depend in a smooth manner on ρ_F and ρ_O. Restricting the focus to generic games Γ (i.e. assuming that payoffs at the end of different branches of the game tree are not identical), we can exclude randomization in period 0.2.
8. The trustors could, for example, assume that only an opportunistic trustee would ever deviate from an equilibrium in which $\rho_F = \rho_O = 1$ by not investing ($\pi^- = 0$). While there is no obvious reason for such pessimism, it would imply that a trustee earns mNP_2 in Γ^- and it would induce the trustees to be willing to establish information sharing even if the cost c is ''very large.''
9. I assume the passive conjecture only for period 0.2. In the sequential equilibrium of the TGs, it is assumed that the trustors are totally convinced that the trustee is of the opportunistic type if there is an unexpected abuse of trust during the phase of the equilibrium in which trust should be placed and honored with certainty (see Kreps and Wilson, 1982a). This reflects that an opportunistic trustee could benefit from such an abuse of trust if the trustors are ''forgiving,'' whereas a friendly trustee would regret the abuse of trust whatever the trustors' reaction is.
10. The effects of changes in T_2 are explained as follows. If $\tau \leq N$, $\bar{c}$ decreases in T_2 because if T_2 is larger, an opportunistic trustee suffers more from having the opportunity to abuse trust only once in Γ^+ rather than m times in Γ^-. On the other hand, if $N < \tau \leq mN$, an opportunistic trustee has a larger incentive to invest if T_2 is larger because he will be trusted and get an opportunity to abuse trust only in Γ^+. If $N < \tau \leq mN$, a friendly trustee's return on investment increases in T_2, too, because if T_2 is larger, the trustors place trust in the randomization phase with higher probability and, hence, expectedly to more honored trust in Γ^+.
11. It can be inferred from the calculations provided in the proof of Proposition 5 that Γ^{ga} could have an equilibrium in which $\rho_F = 1$ and $\rho_O = 0$ if $c_F + T_2 - R_2 \leq c_O$, where c_F and c_O denote the cost of establishing information sharing for friendly and opportunistic trustees, respectively.

12. The comparison of Propositions 3 and 6 shows that there is a gap between the maximum cost $\bar{c}$ for which there are pooling equilibria in which $\rho_F = \rho_O = 1$ and the lower bound on the investment cost for which there are separating equilibria in which $\rho_F = 1$ and $\rho_O = 0$. If the cost c is in that range, the trustee will in equilibrium not establish information sharing. The cost c is so high that the network effects do not warrant the investment and, at the same time, c is so small that it does not deter an opportunistic trustee from investing if he is always trusted if he invests, but never trusted otherwise.

References

Abraham M, Grimm V, Neeß C, et al. (2016) Reputation formation in economic transactions. *Journal of Economic Behavior & Organization* 121(1): 1–14.

Akerlof GA (1970) The market for ''lemons'': quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84(3): 488–500.

Aksoy O and Gambetta D (2016) Behind the veil: the strategic use of religious garb. *European Sociological Review* 32(6): 792–806.

Anderhub V, Engelmann D and Güth W (2002) An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior & Organization* 48(2): 197–216.

Andreoni J (1995) Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics* 110(1): 1–21.

Attanasi G, Battigalli P and Manzoni E (2015) Incomplete-information models of guilt aversion in the trust game. *Management Science* 62(3): 648–667.

Bacharach M and Gambetta D (2001) Trust in signs. In: Cook KS (ed.) *Trust in Society*. New York: Russell Sage, pp. 148–184.

Barclay P (2004) Trustworthiness and competitive altruism can also solve the ''tragedy of the commons.'' *Evolution and Human Behavior* 25(4): 209–220.

Battigalli P and Dufwenberg M (2007) Guilt in games. *American Economic Review* 97(2): 170–176.

Bliege Bird R and Power EA (2015) Prosocial signaling and cooperation among Martu hunters. *Evolution and Human Behavior* 36(5): 389–397.

Bliege Bird R and Smith EA (2005) Signaling theory, strategic interaction, and symbolic capital. *Current Anthropology* 46(2): 221–248.

Bohnet I and Huck S (2004) Repetition and reputation: implications for trust and trustworthiness when institutions change. *American Economic Review* 94(2): 362–366.

Bohnet I, Harmgart H, Huck S, et al. (2005) Learning trust. *Journal of the European Economic Association 3(2–3):* 322–329.

Bolton GE, Katok E and Ockenfels A (2004) How effective are electronic reputation mechanisms? An experimental investigation. *Management Science* 50(11): 1587–1602.

Bower AG, Garber S and Watson JC (1997) Learning about a population of agents and the evolution of trust and cooperation. *International Journal of Industrial Organization* 15(2): 165–190.

Buskens V (2003) Trust in triads: effects of exit, control, and learning. *Games and Economic Behavior* 42(2): 235–252.

Buskens V and Raub W (2002) Embedded trust: control and learning. *Advances in Group Processes* 19: 167–202.

Buskens V and Raub W (2013) Rational choice research on social dilemmas: embeddedness effects on trust. In: Wittek R, Snijders TAB and Nee V (eds) *Handbook of Rational Choice Social Research*. Stanford, CA: Stanford University Press, pp. 113–150.

Buskens V, Frey V and Raub W (2017) Trust games: game-theoretic approaches to embedded trust. In: Uslaner EM (ed.) *Oxford Handbook of Social and Political Trust*. Oxford: Oxford University Press.

Buskens V, Raub W and Van der Veer J (2010) Trust in triads: an experimental study. *Social Networks* 32(4): 301–312.

Camerer CF (2003) *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Camerer CF and Weigelt K (1988) Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56(1): 1–36.

Coleman JS (1990) *Foundations of Social Theory*. Cambridge, MA: Belknap Press of Harvard University Press.

Cook KS and Hardin R (2001) Norms of cooperativeness and networks of trust. In: Hechter M and Opp K-D (eds) *Social Norms*. New York: Russell Sage, pp. 327–347.

Cook KS, Snijders C, Buskens V, et al. (2009) *eTrust: Forming Relationships in the Online World*. New York: Russell Sage.

Cooper D and Kagel J (2013) Other regarding preferences: a selective survey of experimental results. In: Kagel J and Roth A (eds) *Handbook of Experimental Economics, vol*. 2. Princeton, NJ: Princeton University Press, pp. 217–289.

Dasgupta P (1988) Trust as a commodity. In: Gambetta D (ed.) *Trust: Making and Breaking Cooperative Relations*. Oxford: Blackwell, pp. 49–72.

DiMaggio P and Louch H (1998) Socially embedded consumer transactions: for what kinds of purchases do people most often use networks? *American Sociological Review* 63(5): 619–637.

Fehr E and Fischbacher U (2003) The nature of human altruism. *Nature* 425(6960): 785–791.

Fehrler S and Przepiorka W (2013) Charitable giving as a signal of trustworthiness: disentangling the signaling benefits of altruistic acts. *Evolution and Human Behavior* 34(2): 139–145.

Fehrler S and Przepiorka W (2016) Choosing a partner for social exchange: charitable giving as a signal of trustworthiness. *Journal of Economic Behavior & Organization* 129: 157–171.

Feinberg M, Willer R and Schultz M (2014) Gossip and ostracism promote cooperation in groups. *Psychological Science* 25(3): 656–664.

Flap H (2004) Creation and returns of social capital: a new research program. In: Flap H and Völker B (eds) *Creation and Returns of Social Capital*. London: Routledge, pp. 3–23.

Frey V and Van de Rijt A (2016) Arbitrary inequality in reputation systems. *Scientific Reports* 6: 38304.

Frey V, Buskens V and Raub W (2015) Embedding trust: a game theoretic model for investments in and returns on network embeddedness. *Journal of Mathematical Sociology* 39(1): 39–72.

Gambetta D (2009) Signaling. In: Hedström P and Bearman P (eds) *Oxford Handbook of Analytical Sociology*. Oxford: Oxford University Press, pp. 168–194.

Gambetta D and Przepiorka W (2014) Natural and strategic generosity as signals of trustworthiness. *PLoS ONE* 9(5): e97533.

Gintis H, Smith EA and Bowles S (2001) Costly signaling and cooperation. *Journal of Theoretical Biology* 213(1): 103–119.

Guseva A and Rona-Tas A (2001) Uncertainty, risk, and trust: Russian and American credit card markets compared. *American Sociological Review* 66(5): 623–646.

Hillmann H and Aven BL (2011) Fragmented networks and entrepreneurship in late imperial Russia. *American Journal of Sociology* 117(2): 484–538.

Huck S, Lünser GK and Tyran JR (2010) Consumer networks and firm reputation: a first experimental investigation. *Economics Letters* 108(2): 242–244.

James HS (2002) The trust paradox: a survey of economic inquiries into the nature of trust and trustworthiness. *Journal of Economic Behavior & Organization* 47(3): 291–307.

Johnson ND and Mislin AA (2011) Trust games: a meta-analysis. *Journal of Economic Psychology* 32(5): 865–889.

Klein DB (1997) *Reputation: Studies in the Voluntary Elicitation of Good Conduct*. Ann Arbor, MI: University of Michigan Press.

Kollock P (1998) Social dilemmas: the anatomy of cooperation. *Annual Review of Sociology* 24: 183–214.

Kreps DM (1990) *Game Theory and Economic Modeling*. Oxford: Clarendon Press.

Kreps DM and Wilson R (1982a) Reputation and imperfect information. *Journal of Economic Theory* 27(2): 253–279.

Kreps DM and Wilson R (1982b) Sequential equilibria. *Econometrica* 50(4): 863–894.

Kreps DM, Milgrom P, Roberts J, et al. (1982) Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory* 27(2): 245–252.

Lin N (2002) *Social Capital: A Theory of Social Structure and Action*. Cambridge: Cambridge University Press.

Lindenberg S and Steg L (2007) Normative, gain and hedonic goal frames guiding environmental behavior. *Journal of Social Issues* 63(1): 117–137.

Milinski M (2016) Reputation, a universal currency for human social interactions. *Philosophical Transactions of the Royal Society B* 371(1687): 20150100.

Milinski M, Semmann D and Krambeck H (2002) Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society of London—Series B: Biological Sciences* 269(1494): 881–883.

Paik A and Woodley V (2012) Symbols and investments as signals: courtship behaviors in adolescent sexual relationships. *Rationality and Society* 24(1): 3–36.

Patel DS (2012) Concealing to reveal: the informational role of Islamic dress. *Rationality and Society* 24(3): 295–323.

Prendergast C (1999) The provision of incentives in firms. *Journal of Economic Literature* 37(1): 7–63.

Przepiorka W (2013) Buyers pay for and sellers invest in a good reputation: more evidence from Ebay. *Journal of Socio-Economics* 42: 31–42.

Przepiorka W and Berger J (2017) Signaling theory evolving: signals and signs of trustworthiness in social exchange. In: Jann B and Przepiorka W (eds) *Social Dilemmas, Institutions and the Evolution of Cooperation*. Berlin: De Gruyter Oldenbourg, pp. 373–392.

Przepiorka W and Diekmann A (2013) Temporal embeddedness and signals of trustworthiness: experimental tests of a game theoretic model in the United Kingdom, Russia, and Switzerland. *European Sociological Review* 29(5): 1010–1023.

Rasmusen E (1994) *Games and Information: An Introduction to Game Theory*. 4th ed. Oxford: Blackwell.

Raub W (2004) Hostage posting as a mechanism of trust: binding, compensation, and signaling. *Rationality and Society* 16(3): 319–365.

Raub W, Buskens V and Frey V (2013) The rationality of social structure: cooperation in social dilemmas through investments in and returns on social capital. *Social Networks* 35(4): 720–732.

Resnick P, Kuwabara K, Zeckhauser R, et al. (2000) Reputation systems. *Communications of the ACM* 43(12): 45–48.

Rubinstein A (1985) Choices of conjectures in a bargaining game with incomplete information. In: Roth AE (ed.) *Game-Theoretic Models of Bargaining*. Cambridge: Cambridge University Press, pp. 99–114.

Snijders C (1996) *Trust and Commitments*. Amsterdam: Thesis Publishers.

Snijders C and Weesie J (2009) Online programming markets. In: Cook KS, Snijders C, Buskens V, et al. (eds) *Etrust: Forming Relationships in the Online World*. New York: Russell Sage, pp. 166–186.

Sosis R (2005) Does religion promote trust? The role of signaling, reputation, and punishment. *Interdisciplinary Journal of Research on Religion* 1: 1–30.

Spence M (1973) Job market signaling. *Quarterly Journal of Economics* 87(3): 355–374.

Zahavi A (1975) Mate selection: a selection for a handicap. *Journal of Theoretical Biology* 53(1): 205–214.

Appendix 1

This appendix provides a lemma that specifies the equilibria of Γ^- and Γ^+ formally and it provides the proofs for Propositions 1–6.

Equilibrium behavior in Γ^- and Γ^+

Lemma 1 specifies the sequential equilibrium of Γ^- and Γ^+. Lemma 1 uses notation that accounts for the difference between Γ^- and Γ^+ and it applies to both continuation games simultaneously. f denotes the number of remaining periods in which the trustor at play will have information about the current period. In Γ^+, f is simply the total number of remaining periods. In Γ^-, f is the number of TGs that the current trustor will still play after the current period. Period "-1" is used to indicate the last period on which the trustor has information. For Γ^+, period -1 refers simply to the immediately preceding period. For Γ^-, period -1 refers to the last period in which the current trustor was at play. Note that Lemma 1 specifies the belief of a trustor only for the case that there is a period -1. If there is no period -1, a trustor's belief equals the belief entering the continuation game, namely, π^- or π^+.

Lemma 1. *The beliefs and strategies specified below constitute a sequential equilibrium in Γ^- and Γ^+.*

- *Belief of trustor i in period n that the trustee is of the friendly type*:
  - *If in period -1 trust was not placed, then $\pi^i_n = \pi^i_{-1}$.*
  - *If in period -1 trust was placed and honored, then $\pi^i_n = \max(RISK^{f+1}, \pi^i_{-1})$.*
  - *If in period -1 trust was placed and abused, then $\pi^i_n = 0$.*
- *Probability that (if at play) trustor i places trust in period n*:
  - *If $\pi^i_n > RISK^{f+1}$, then $t^i_n = 1$.*
  - *If $\pi^i_n = RISK^{f+1}$, then $t^i_n = TEMP$.*
  - *If $\pi^i_n < RISK^{f+1}$, then $t^i_n = 0$.*
- *Probability that an opportunistic trustee honors trust of the trustor i at play in period n*:
  - *If $\pi^i_n \geq RISK^f$, then $h_n = 1$.*
  - *If $\pi^i_n < RISK^f$, then $h_n = (\pi^i_n / 1 - \pi^i_n)\left((1/RISK^f) - 1\right)$.*

Lemma 1 is a generalization of Theorems 1 and 3 of Frey et al. (2015) and I refer the reader to Frey et al. (2015) for the proof.

Proof of Propositions 1 and 2: payoffs in Γ^- and Γ^+

Proposition 1 is implied by the sequential equilibrium of the TGs in Γ^-, which is described in section ''Behavior and payoffs in the TGs'' and specified formally in Lemma 1. For the case that $\tau^- \leq N$, a trustee's expected payoff for the TGs with each trustor i in Γ^- can be thought of as consisting of two main components. The first component pertains to the first phase of the equilibrium in which trust is placed and honored with certainty. In Γ^{ga}, this component is $(N - \tau^-)R_2$ for either type of trustee. In Γ^{rs}, this component is likewise $(N - \tau^-)R_2$ for an opportunistic trustee while it is $(N - \tau^-)(R_2 + \upsilon)$ for a friendly trustee.

The second component pertains to the τ^- last TGs that the trustee plays with trustor i (the second and third phase of the equilibrium). Assume, first, that the trustee is of the friendly type. In this case, trust may break down from the second of these τ^- last TGs on because trustor i begins to randomize, placing trust with probability *TEMP* as long as she placed trust before. This leads to an expected payoff for a friendly trustee for the τ^- last TGs played with trustor i of

$$\sum_{i=0}^{\tau-1} \left(TEMP^i \cdot R_2 + (1 - TEMP^i)P_2\right) \tag{2}$$

in Γ^{ga} and

$$\sum_{i=0}^{\tau-1} \left(TEMP^i \cdot (R_2 + \upsilon) + (1 - TEMP^i)P_2\right) \tag{3}$$

in Γ^{rs}. Using $\sum_{i=0}^{n} x^i = (1 - x^{n+1})/(1 - x)$, equations (2) and (3) can be rearranged to $\tau^- P_2 + (T_2 - P_2)(1 - TEMP^{\tau^-})$ and $\tau^- P_2 + (T_2 - P_2)\left((R_2 + \upsilon - P_2)/(R_2 - P_2)\right)(1 - TEMP^{\tau^-})$, respectively.

Now assume that the trustee is of the opportunistic type. In the first of the τ^- last TGs that he plays with trustor i, the opportunistic trustee is indifferent between, on one hand, honoring trust and maybe being trusted again by trustor i and, on the other hand, abusing trust and certainly never being trusted again by trustor i. In the latter case, the trustee earns $T_2 + (\tau^- - 1)P_2$ and this is thus his expected payoff for the τ^- last TGs with trustor i.

Proposition 2 is implied by the sequential equilibrium described in section ''Behavior and payoffs in the TGs'' and specified formally in Lemma 1 in the same manner as Proposition 1 is implied by the sequential equilibrium of the interactions between the trustee and one trustor i in Γ^-. In Γ^+, it is as

if the trustee played mN TGs with one single trustor. So, the first component of this expected payoff becomes $(mN - \tau)R_2$ (for an opportunistic or guilt-avoiding trustee) while the second component remains unchanged.

Proof of Proposition 3: the condition for the existence of equilibria in which $\rho_F = \rho_O = 1$

In general, Γ has a sequential equilibrium in which $\rho_F = \rho_O = 1$ if $c \leq \bar{c} = \min(U_F^{\Gamma^+} - U_F^{\Gamma^-}, U_O^{\Gamma^+} - U_O^{\Gamma^-})$. Otherwise, if $c > \bar{c}$, at least one type of trustee would be better off if he deviates from the conjectured equilibrium by not investing. To specify $\bar{c}$, recall that if $\rho_F = \rho_O = 1$, Bayes' rule implies that $\pi^+ = \pi$. Furthermore, the passive conjecture refinement requires that the trustors' out-of-equilibrium belief π^- likewise equals π. So, in an equilibrium in which $\rho_F = \rho_O = 1$, $\pi^- = \pi^+ = \pi$ and, hence, $\tau^- = \tau^+$. Let us again use τ to denote τ^- and τ^+ simultaneously and use, for example, $U_F^{\Gamma^+(\tau)}$ to denote the expected payoff of a friendly trustee for Γ^+ given the τ (i.e. τ^+) that results from $\pi^+ = \pi$. We can then formulate the condition for the existence of an equilibrium in which $\rho_F = \rho_O = 1$ as $c \leq \bar{c} = \min(U_F^{\Gamma^+(\tau)} - U_F^{\Gamma^-(\tau)}, U_O^{\Gamma^+(\tau)} - U_O^{\Gamma^-(\tau)})$.

Consider first Γ^{ga}. To calculate $U_F^{\Gamma^{ga+}(\tau)} - U_F^{\Gamma^{ga-}(\tau)}$ and $U_O^{\Gamma^{ga+}(\tau)} - U_O^{\Gamma^{ga-}(\tau)}$, we can use the formulas for expected payoffs in Propositions 1 and 2. For $\tau \leq N$, we obtain

$$\begin{aligned} U_F^{\Gamma^{ga+}(\tau)} - U_F^{\Gamma^{ga-}(\tau)} &= (m-1)\big(\tau(R_2 - P_2) - (T_2 - P_2)(1 - TEMP^\tau)\big), \\ U_O^{\Gamma^{ga+}(\tau)} - U_O^{\Gamma^{ga-}(\tau)} &= (m-1)\big(\tau(R_2 - P_2) - (T_2 - P_2)\big). \end{aligned} \tag{4}$$

As $0 < \text{TEMP} < 1$, it holds that $U_F^{\Gamma^{ga+}(\tau)} - U_F^{\Gamma^{ga-}(\tau)} > U_O^{\Gamma^{ga+}(\tau)} - U_O^{\Gamma^{ga-}(\tau)}$ and, hence, $\bar{c}^{ga} = (m-1)\big(\tau(R_2 - P_2) - (T_2 - P_2)\big)$ if $\tau \leq N$.

For $N < \tau \leq mN$, we have

$$\begin{aligned} U_F^{\Gamma^{ga+}(\tau)} - U_F^{\Gamma^{ga-}(\tau)} &= (mN - \tau)(R_2 - P_2) + (T_2 - P_2)(1 - TEMP^\tau) \\ U_O^{\Gamma^{ga+}(\tau)} - U_O^{\Gamma^{ga-}(\tau)} &= (mN - \tau)(R_2 - P_2) + (T_2 - P_2) \end{aligned} \tag{5}$$

where $U_F^{\Gamma^{ga+}(\tau)} - U_F^{\Gamma^{ga-}(\tau)} < U_O^{\Gamma^{ga+}(\tau)} - U_O^{\Gamma^{ga-}(\tau)}$ and thus $\bar{c}^{ga} = (mN - \tau)(R_2 - P_2) + (T_2 - P_2)(1 - TEMP^\tau)$ if $N < \tau \leq mN$.

Now consider Γ^{rs}. For $\tau \leq N$, it is useful to introduce an alternative way of expressing the trustees' expected payoffs for Γ^{rs-} and Γ^{rs+}.

Alternative to the formulas in Propositions 1 and 2, we express these payoffs for $\tau \leq N$ as follows:

$$U_F^{\Gamma^{rs-}} = m\left((N-\tau^-)(R_2+\upsilon) + \sum_{i=0}^{\tau^- -1}\left(TEMP^i(R_2+\upsilon) + (1-TEMP^i)P_2\right)\right) \tag{6a}$$

$$U_F^{\Gamma^{rs+}} = (mN-\tau^+)(R_2+\upsilon) + \sum_{i=0}^{\tau^+ -1}\left(TEMP^i(R_2+\upsilon) + (1-TEMP^i)P_2\right) \tag{6b}$$

$$U_O^{\Gamma^{rs-}} = m\left((N-\tau^-)R_2 + \sum_{i=0}^{\tau^- -2}\left(TEMP^i R_2 + (1-TEMP^i)P_2\right) + TEMP^{\tau^- -1}T_2 + (1-TEMP^{\tau^- -1})P_2\right) \tag{6c}$$

$$U_O^{\Gamma^{rs+}} = (mN-\tau^+)R_2 + \sum_{i=0}^{\tau^+ -2}\left(TEMP^i R_2 + (1-TEMP^i)P_2\right) + TEMP^{\tau^+ -1}T_2 + (1-TEMP^{\tau^+ -1})P_2 \tag{6d}$$

Equations (6a) and (6b) for the friendly trustee are directly implied by the sequential equilibrium in the way explained in the proof of Propositions 1 and 2. It is only that we do here *not* rearrange the summation. Equations (6c) and (6d) for the opportunistic trustee are obtained when assuming that the trustee does—if trust gets placed—honor trust in the first $\tau - 1$ TGs after the start of the randomization phase and abuses trust in the very last TG if trust is still placed in that TG. In the first of these TGs, trust gets placed with certainty and from the second of these TGs on, the trustor(s) place trust with probability TEMP as long as trust was always placed before. From equations (6a) and (6b), we obtain

$$U_F^{\Gamma^{rs+}(\tau)} - U_F^{\Gamma^{rs-}(\tau)} = (m-1)\left(\tau - \sum_{i=0}^{\tau-1} TEMP^i\right)(R_2+\upsilon-P_2) \tag{7}$$

From equations (6c) and (6d), we obtain

$$U_O^{\Gamma^{rs+}(\tau)} - U_O^{\Gamma^{rs-}(\tau)} = (m-1)\left(\tau - \sum_{i=0}^{\tau-1} TEMP^i\right)(R_2 - P_2) - TEMP^{\tau-1}(T_2 - R_2) \tag{8}$$

The comparison of equations (7) and (8) shows that $U_F^{\Gamma^{rs+}(\tau)} - U_F^{\Gamma^{rs-}(\tau)} > U_O^{\Gamma^{rs+}(\tau)} - U_O^{\Gamma^{rs-}(\tau)}$ because $(R_2 + \upsilon - P_2) > (R_2 - P_2)$ and $TEMP^{\tau-1}(T_2 - R_2) > 0$. Thus, if $\tau \leq N$, $\bar{c}^{rs} = U_O^{\Gamma^{rs+}(\tau)} - U_O^{\Gamma^{rs-}(\tau)}$. The specification of $\bar{c}^{rs}$ provided in Proposition 3, namely, $\bar{c}^{rs} = (m-1)\big(\tau(R_2 - P_2) - (T_2 - P_2)\big)$, follows from the expressions for $U_O^{\Gamma^{rs-}}$ and $U_O^{\Gamma^{rs+}}$ in Propositions 1 and 2 and can also be obtained from equation (8).

For the scenario that $N < \tau \leq mN$, we derive $U_F^{\Gamma^{rs+}(\tau)} - U_F^{\Gamma^{rs-}(\tau)}$ using the same manner of expressing expected payoffs as in equations (6a) and (6b). We have

$$\begin{aligned} U_F^{\Gamma^{rs+}(\tau)} - U_F^{\Gamma^{rs-}(\tau)} &= (mN - \tau)(R_2 + \upsilon) \\ &\quad + \sum_{i=0}^{\tau-1} \left(TEMP^i(R_2 + \upsilon) + (1 - TEMP^i)P_2\right) - mNP_2 \\ &= \left(mN - \tau + \sum_{i=0}^{\tau-1} TEMP^i\right)(R_2 + \upsilon - P_2) \end{aligned} \tag{9}$$

For the opportunistic trustee, we obtain from Propositions 1 and 2 that if $N < \tau \leq mN$

$$U_O^{\Gamma^{rs+}(\tau)} - U_O^{\Gamma^{rs-}(\tau)} = (mN - \tau)(R_2 - P_2) + (T_2 - P_2) \tag{10}$$

It follows from equations (9) and (10) that $U_F^{\Gamma^{rs+}(\tau)} - U_F^{\Gamma^{rs-}(\tau)} > U_O^{\Gamma^{rs+}(\tau)} - U_O^{\Gamma^{rs-}(\tau)}$. Specifically, the inequality $U_F^{\Gamma^{rs+}(\tau)} - U_F^{\Gamma^{rs-}(\tau)} > U_O^{\Gamma^{rs+}(\tau)} - U_O^{\Gamma^{rs-}(\tau)}$ can be reduced to

$$(mN - \tau)\upsilon + \sum_{i=0}^{\tau-1} TEMP^i(R_2 + \upsilon - P_2) > (T_2 - P_2) \tag{11}$$

Equation (11) must hold because $(R_2 + \upsilon - P_2) > (T_2 - P_2)$ and $\sum_{i=0}^{\tau-1} TEMP^i \geq 1$. This shows that $\bar{c}^{rs} = U_O^{\Gamma^{rs+}(\tau)} - U_O^{\Gamma^{rs-}(\tau)} = (mN - \tau)(R_2 - P_2) + (T_2 - P_2)$ if $N < \tau \leq mN$.

Proof of Proposition 4: comparative statics of the condition for the existence of an equilibrium with $\rho_F = \rho_O = 1$

For the case that $N < \tau \leq mN$, it will be useful to have an expression for $\bar{c}^{ga} = U_F^{\Gamma^{ga+}(\tau)} - U_F^{\Gamma^{ga-}(\tau)}$ derived using a formulation of $U_F^{\Gamma^{ga+}}$ that is equivalent to the formulation of $U_F^{\Gamma^{rs+}}$ in equation (6b). For $N < \tau \leq mN$, we then have

$$\begin{aligned} \bar{c}^{ga} &= (mN - \tau)R + \sum_{i=0}^{\tau-1} \left(TEMP^i \cdot R + (1 - TEMP^i)P\right) - mNP_2 \\ &= \left(mN - \tau + \sum_{i=0}^{\tau-1} TEMP^i\right)(R_2 - P_2) \end{aligned} \tag{12}$$

Effects of changes in π and the trustors' payoffs in the TG. The trustors' payoffs in the TG and the probability π affect $\bar{c}^{ga}$ and $\bar{c}^{rs}$ exclusively through τ. For the case that $\tau \leq N$, it follows directly from the expressions in Proposition 3, namley, $\bar{c}^{ga} = \bar{c}^{rs} = (m-1)\left(\tau(R_2 - P_2) - (T_2 - P_2)\right)$, that $\bar{c}^{ga}$ and $\bar{c}^{rs}$ increase by $(m-1)(R_2 - P_2)$ for every unit increase in τ. $\tau = \lceil \log(\pi)/\log(RISK) \rceil$ decreases stepwise in π and increases stepwise in $RISK$. $RISK = (P_1 - S_1)/(R_1 - S_1)$, in turn, increases in P_1 and decreases in R_1 and S_1 (specifically, $\partial RISK/\partial P_1 = 1/(R_1 - S_1) > 0$, $\partial RISK/\partial R_1 = -(P_1 - S_1)/(R_1 - S_1)^2 < 0$, and $\partial RISK/\partial S_1 = -(R_1 - P_1)/(R_1 - S_1)^2 < 0$). Hence, if a decrease in π or an increase in $RISK$ (caused by an increase in P_1 or a decrease in R_1 or S_1) leads to an increase in τ, $\bar{c}^{ga}$ and $\bar{c}^{rs}$ increase if $\tau \leq N$.

For $N < \tau \leq mN$, we need to consider Γ^{ga} and Γ^{rs} separately. For Γ^{ga}, it follows from equation (12) that an increase in τ leads to a decrease in $\bar{c}^{ga}$. An increase in τ by 1 implies that "one additional $R_2 - P_2$" is subtracted while less than "one additional $R_2 - P_2$" is added in the summation (since $0 < TEMP < 1$). For Γ^{rs}, it follows from the expression in Proposition 3, namely, $\bar{c}^{rs} = (mN - \tau)(R_2 - P_2) + (T_2 - P_2)$, that if $N < \tau \leq mN$, $\bar{c}^{rs}$ decreases by $R_2 - P_2$ for every unit increase in τ. Hence, if a decrease in π or an increase in $RISK$ (caused by an increase in P_1 or a decrease in R_1 or S_1) leads to an increase in τ, $\bar{c}^{ga}$ as well as $\bar{c}^{rs}$ decrease if $N < \tau \leq mN$.

Effects of changes in the trustee's payoffs in the TG. Consider, first, an increase in T_2. It is easy to see that if $\tau \leq N$, $\bar{c}^{ga} = \bar{c}^{rs} = (m-1)\left(\tau(R_2 - P_2) - (T_2 - P_2)\right)$ decreases in T_2. For $N < \tau \leq mN$, the effect of a change in T_2 on $\bar{c}^{ga}$ can be inferred from equation (15). If T_2 increases, $TEMP = (T_2 - R_2)/(T_2 - P_2)$ increases $\left(\partial TEMP/\partial T_2 = (R_2 -\right.$

$P_2)/(T_2 - P_2)^2 > 0\big)$. Hence, if $N < \tau \leq mN$, $\bar{c}^{ga}$ increases in T_2 because if T_2 is larger, every element that is added in the summation is larger. That, given $N < \tau \leq mN$, $\bar{c}^{rs}$ likewise increases in T_2 follows directly from the respective expression in Proposition 3, namely, $\bar{c}^{rs} = (mN - \tau)(R_2 - P_2) + (T_2 - P_2)$.

Now consider an increase in R_2. For $\tau \leq N$, it follows straightforwardly from $\bar{c}^{ga} = \bar{c}^{rs} = (m-1)\big(\tau(R_2 - P_2) - (T_2 - P_2)\big)$ that $\bar{c}^{ga}$ and $\bar{c}^{rs}$ increase in R_2. How $\bar{c}^{ga}$ changes in R_2 if $N < \tau \leq mN$ follows from taking the derivative of $\bar{c}^{ga} = (mN - \tau)(R_2 - P_2) + (T_2 - P_2)(1 - TEMP^{\tau})$. This gives $\partial\bar{c}^{ga}/\partial R_2 = mN - \tau(1 - TEMP^{\tau-1})$, which, given $\tau \leq mN$, is larger than 0. For Γ^{rs}, we obtain from $\bar{c}^{rs} = (mN - \tau)(R_2 - P_2) + (T_2 - P_2)$ that, given $N < \tau \leq mN$, $\partial\bar{c}^{rs}/\partial R_2 = mN - \tau$. This shows that $\bar{c}^{rs}$ increases in R_2 if $N < \tau < mN$ and does not change in R_2 if $\tau = mN$.

Now consider an increase in P_2. For $\tau \leq N$, we obtain from $\bar{c}^{ga} = \bar{c}^{rs} = (m-1)\big(\tau(R_2 - P_2) - (T_2 - P_2)\big)$ that $\partial\bar{c}^{ga}/\partial P_2 = \partial\bar{c}^{rs}/\partial P_2 = (m-1)(1-\tau)$. This shows that $\bar{c}^{ga}$ and $\bar{c}^{rs}$ decrease in P_2 if $\tau > 1$, while they are independent of P_2 if $\tau = 1$. For $N < \tau \leq mN$, we obtain from $\bar{c}^{ga} = (mN - \tau)(R_2 - P_2) + (T_2 - P_2)(1 - TEMP^{\tau})$ that $\partial\bar{c}^{ga}/\partial P_2 = (\tau - 1)(1 - TEMP) - mN$, which, given $\tau \leq mN$, is smaller than 0. For Γ^{rs}, we derive from $\bar{c}^{rs} = (mN - \tau)(R_2 - P_2) + (T_2 - P_2)$ that $\partial\bar{c}^{rs}/\partial P_2 = \tau - 1 - mN$, which given $\tau \leq mN$ is likewise smaller than 0. This proves that $\bar{c}^{ga}$ and $\bar{c}^{rs}$ decrease in P_2 if $1 < \tau \leq mN$.

Effects of changes in m and N. For changes in m we have $\partial\bar{c}^{ga}/\partial m = \partial\bar{c}^{rs}/\partial m = \tau(R_2 - P_2) - (T_2 - P_2)$ if $\tau \leq N$. $\tau(R_2 - P_2)$ is not necessarily larger than $T_2 - P_2$ and, therefore, it is unclear whether an increase in m leads to an increase or a decrease in $\bar{c}$. If $N < \tau \leq mN$, $\partial\bar{c}^{ga}/\partial m = \partial\bar{c}^{rs}/\partial m = N(R_2 - P_2)$, which shows that $\bar{c}$ increases in m if $N < \tau \leq mN$.

Changes in N also affect $\bar{c}$ only if $N < \tau \leq mN$. If $\tau \leq N$, $\partial\bar{c}^{ga}/\partial N = \partial\bar{c}^{rs}/\partial N = 0$. If $N < \tau \leq mN$, $\partial\bar{c}^{ga}/\partial N = \partial\bar{c}^{rs}/\partial N = m(R_2 - P_2) > 0$.

Proof of Proposition 5: the impossibility of an equilibrium in which $\rho_F = 1$ and $\rho_O = 0$ in Γ^{ga}

The combination $\rho_F = 1$ and $\rho_O = 0$ implies (by Bayes' rule) $\pi^- = 0$ and $\pi^+ = 1$ and, consequently, $\tau^- = \infty$ and $\tau^+ = 1$. Given this, a friendly trustee will indeed want to play $\rho_F = 1$ if $c \leq U_F^{\Gamma^{ga+}(\tau^+ = 1)} - U_F^{\Gamma^{ga-}(\tau^- = \infty)} = mN(R_2 - P_2)$. But if this holds, an opportunistic trustee will want to play

$\rho_O = 1$, too, because if $c \leq mN(R_2 - P_2)$, it also holds that $c < U_O^{\Gamma^{ga+}(\tau^+=1)} - U_O^{\Gamma^{ga-}(\tau^-=\infty)} = mN(R_2 - P_2) + T_2 - R_2$. Consequently, the combination $\rho_F = 1$ and $\rho_O = 0$ cannot be part of a sequential equilibrium of Γ^{ga}.

Proof of Proposition 6: the condition for the existence of an equilibrium in which $\rho_F = 1$ and $\rho_O = 0$ in Γ^{rs}

Again, the combination $\rho_F = 1$ and $\rho_O = 0$ implies $\pi^- = 0$ and $\pi^+ = 1$ and, hence, $\tau^- = \infty$ and $\tau^+ = 1$. Given $\tau^- = \infty$ and $\tau^+ = 1$, $\rho_F = 1$ can be part of an equilibrium strategy for a friendly trustee if $c \leq U_F^{\Gamma^{rs+}(\tau^+=1)} - U_F^{\Gamma^{rs-}(\tau^-=\infty)} = mN(R_2 + \upsilon) - mNP_2$, while $\rho_O = 0$ can be part of an equilibrium strategy for an opportunistic trustee if $c \geq U_O^{\Gamma^{rs+}(\tau^+=1)} - U_O^{\Gamma^{rs-}(\tau^-=\infty)} = (mN - 1)R_2 + T_2 - mNP_2$. Hence, the combination $\rho_F = 1$ and $\rho_O = 0$ is part of a sequential equilibrium of Γ^{rs} if $mN(R_2 - P_2) + T_2 - R_2 \leq c \leq mN(R_2 + \upsilon - P_2)$. That there exists some cost c for which this holds is implied by the assumption that $T_2 < R_2 + \upsilon$.