TRUST AND TRUSTWORTHINESS IN AN ECONOMY WITH HETEROGENEOUS INDIVIDUALS

Peter Katuščák
Joel Slemrod

CERGE-EI

Charles University
Center for Economic Research and Graduate Education
Academy of Sciences of the Czech Republic
Economics Institute

WORKING PAPER SERIES (ISSN 1211-3298)
Electronic Version

305

Working Paper Series **305**
(ISSN 1211-3298)

Trust and Trustworthiness in an Economy with Heterogeneous Individuals

Peter Katuščák
Joel Slemrod

CERGE-EI
Prague, September 2006

ISBN 80-7343-101-7 (Univerzita Karlova. Centrum pro ekonomický výzkum a doktorské studium)
ISBN 80-7344-090-3 (Akademie věd České republiky. Národohospodářský ústav)

Trust and Trustworthiness in an Economy with Heterogeneous Individuals *

Peter Katuščák† Joel Slemrod‡

September 2006

Abstract

We analyze the determinants of trust and trustworthiness in a matching equilibrium when agents have heterogeneous predispositions towards trusting and trustworthy behavior, there is transmission of information via both individual and collective reputations, and successful matches may persist. In new matches, more social trustworthiness breeds more individual trust. However, whether more social trust breeds more or less individual trustworthiness depends on the observability of individual histories of play. If it is low, more trust generally breeds less trustworthiness, while if it is high, more trust breeds more trustworthiness. We combine the links between social trust and trustworthiness to construct a general trust/trustworthiness equilibrium and discuss its properties.

Keywords: trust, trustworthiness, reputation.
JEL Classification: C7

*We would like to thank Andreas Ortmann, Avner Shaked, Richard Stock, seminar participants at the University of Michigan, and conference participants at the 2003 PET Meetings for helpful comments and suggestions. Any remaining errors are our own.

†CERGE-EI, P.O.Box 882, Politických vězňů 7, 111 21 Praha 1, Czech Republic, Peter.Katuscak@cerge-ei.cz. CERGE-EI is a joint workplace of the Center for Economic Research and Graduate Education, Charles University, and the Economics Institute of the Academy of Sciences of the Czech Republic.

‡Department of Economics, University of Michigan, 611 Tappan Street, Ann Arbor, MI 48109-1220, jslemrod@umich.edu.

Abstrakt

Táto štúdia analyzuje faktory ovplyvňujúce dôveru a dôveryhodnosť v rovnovážnom stave, keď ekonomickí agenti majú heterogénne predispozície pre dôverné a dôveryhodné správanie, informácie sa prenášajú prostredníctvom individuálnych a kolektívnych reputácií, a úspešné transakčné vzťahy môžu pretrvávať. Pri nových vzťahoch, viac spoločenskej dôveryhodnosti prináša viac individuálnej dôvery. Avšak či viac spoločenskej dôvery prináša viac alebo menej individuálnej dôveryhodnosti závisí na tom, do akej miery ekonomickí agenti poznajú správanie sa svojich potenciálnych partnerov v minulosti. Keď je táto znalosť malá, viac spoločenskej dôvery vedie k menšej individuálnej dôveryhodnosti, ale keď je táto znalosť veľká, viac spoločenskej dôvery vedie k väčšej individuálnej dôveryhodnosti. Táto štúdia potom spája tieto dve závislosti medzi spoločenskou dôverou a dôveryhodnosťou a analyzuje celkový rovnovážny stav a jeho vlastnosti.

1 Introduction

The notions of trust and trustworthiness have received much recent attention in social science, stimulated in part by the work of Putnam (1993) and Fukuyama (1995), but with antecedents in, for example, Coleman (1990). Economists have for a long time recognized the critical role played by trust in economic performance. Arrow, for example, remarks: "Virtually every commercial transaction has within itself an element of trust, certainly any transaction conducted over a period of time. It can plausibly be argued that much of the economic backwardness in the world can be explained by the lack of mutual confidence" (1972, p. 357). In high-trust societies, individuals need to spend less resources to protect themselves from being exploited in economic transactions. Knack and Keefer (1997) provide evidence that trusting societies tend to have stronger incentives to innovate and to accumulate both physical and human capital and, as a result, grow faster, and Zak and Knack (2001) corroborate the positive effect of aggregate trust on growth.

The flip side of trust is trustworthiness. Whereas trust can be defined as the commitment of resources to an activity where the outcome depends upon the cooperative behavior of others, trustworthiness can be defined as behavior that increases the returns to people who display trust toward the person. The idea of reputation–the level of trust one is perceived to merit–has also been examined. As Axelrod (1986) puts it, an individual's reputation derives from the adherence to or violation of a norm that others view as a signal about the individual's future behavior in a wide variety of situations.

Whether an individual trusts a potential business partner has traditionally been modeled in the economic literature as a matter of the partner's reputation for his type, or, more precisely, a belief about the partner's type when this type is imperfectly observed. One strand of literature, represented by Sobel (1985), Watson (1999), and Blonski and Probst (2001), analyzes the formation of reputation in repeated games with a fixed set of players. These authors show that mutual trust builds up over time as partners start by commit-

ting small amounts of resources early in the game to "get to know" their opponents, and successful experiences lead to an increase in the scale of cooperation over time.

However, most realistic situations involve games in which opponents may change over time. Once this is acknowledged, two distinct new questions arise. First, how likely is it that information about individual histories of play gets transmitted from one match to another? Second, how long do individual matches last, and how often do the partners change? Or, in other words, what is the relative importance of repeated matches versus rematching? Because of the changing partner character of the game, the literature on the topic, including this paper, utilizes the random matching framework pioneered by Rosenthal (1979). In this framework, individuals meet randomly in any given period to form potentially mutually beneficial matches. In Rosenthal's original work, all the matches last for one period, and hence there is no role for the continuation of cooperation over time. Tirole (1996), using this framework, considers varying degrees of observability of individual histories of play, which leads players to utilize both individual and collective reputations when forming their beliefs about their opponents' types. Ghosh and Ray (1996) extend Rosenthal's framework by allowing for repeated interaction within a given match. This repeated interaction leads to the buildup of mutual trust over time as in a repeated game with a fixed set of players. However, unlike Tirole, they do not allow for any observability of individual reputations from one match to another.[1]

In addition, trust or trustworthiness may not be entirely calculative, i.e., they may be based on other factors beside the partner's reputation and the personal gain from cheating vs. the effect on one's own reputation. For example, Alesina and La Ferrara (2002) identify a number of social factors driving trust, such as a recent traumatic experience or a certain ethnic/racial background, that are not necessarily related to the reputation of

[1]There are several other recent theoretical contributions addressing the issue of trust. For example, Dixit (2003) considers the role of individual reputations in a random matching model and studies an endogenous process of the rise of informational intermediaries that track these reputations. Chen (2000) develops a static model in which individuals differ in their intrinsic preferences for being honest, or trustworthy, and uses the model to analyze the role of trust in contracting.

current potential transaction partners. There are also recent experiments by Ashraf et al. (2006) conducted in Russia, South Africa, and the U.S. that suggest that even though expectations of return, or partner's reputation, account for most of the interpersonal variance in trust, "unconditional kindness" matters too, and the same is true of trustworthiness. It is therefore likely that individual trust and trustworthiness are driven by factors that are unrelated to the material payoffs involved in current and future potential transactions and that are heterogeneous across individuals.

Based on these motivations, the current paper extends the previous theoretical literature in three directions. First, we provide a unifying framework by developing a general model that allows both an arbitrary degree of observability of individual histories of play from one match to another, and an arbitrary durability of (successful) individual matches over time. Second, our model features an arbitrary intensity of the matching process, which generally leads to a subset of players who are unmatched in a given period. Third, in order to capture "unconditional kindness," we allow heterogeneity in the predispositions for trusting and trustworthy behavior, which leads to a sorting of individuals into trusting and mistrusting on the one hand, and into trustworthy and untrustworthy on the other hand. This is in contrast to the heterogeneity considered in Ghosh and Ray (1996), who do not separate trusting and trustworthy behavior, and the heterogeneity considered by Tirole (1996), who only considers differences in attitudes towards trustworthiness. In addition, rather than separating players into "rational" (i.e., those who maximize their payoff) and "dogmatic" (i.e., those who always follow some prescribed strategy), we assume that there is a disutility associated with not trusting and a disutility associated with being untrustworthy, and that both of these behavioral predispositions have a continuous distribution in the population. Therefore, rather than having dogmatic players following their prescribed course of action and rational players following a utility maximizing action, the players in our model continuously sort between trusting and not trusting and between being trustworthy and being untrustworthy based on strategic considerations and

their individual behavioral predispositions.

We analyze trust and trustworthiness in an infinitely repeated random matching environment where the stage game is similar to the one analyzed by Berg et al. (1995). In the stage game that we consider, the first mover, called the initiator, has the option to initiate or not initiate a transaction. Initiating a transaction involves the commitment of a certain amount of (investment) resources that may potentially be stolen. If the transaction is initiated, the second mover, called the respondent, has the option to respond honestly or dishonestly. In the case of an honest response, both players gain. In the case of a dishonest response, the respondent simply absconds with the resources put forward by the initiator. In this game, initiating corresponds to trust, while not initiating corresponds to mistrust. Similarly, responding honestly corresponds to trustworthiness, while responding dishonestly corresponds to a lack of trustworthiness. Apart from the pecuniary payoffs, the players' utilities are also affected by their behavioral predispositions. In particular, each initiator has a certain disutility from not trusting, and each respondent has a certain disutility from being untrustworthy. Therefore the extent of trusting and trustworthy behavior in the stage game is affected by both the structure of the pecuniary payoffs and these behavioral predispositions.

In the random matching environment, there are both matched and unmatched initiators and respondents at the beginning of each period. A subset of each are randomly matched into pairs (using uniform matching), with each pair consisting of one initiator and one respondent. The players in both the pre-existing and the new matches then play the investment game described earlier. Successful matches (i.e., matches in which the initiators and the respondents exhibit trust and trustworthiness, respectively) have a positive probability of survival until the next period. However, a fraction of these matches do not survive and their participants enter the pool of unmatched players for the next period. All other matches dissolve immediately and their participants enter the pool of unmatched players for the next period.

Assuming that a particular respondent's net gain (after disutility from being dishonest) from behaving dishonestly exceeds the gain from behaving honestly, the unique subgame perfect equilibrium for the one-shot stage game is for the respondent to respond dishonestly and, consequently, for the initiator not to initiate. Given that a match breaks up once there is a dishonest response, honesty can only be induced if dishonest respondents get punished by their future opponents. In particular, if an initiator observes that the respondent behaved dishonestly in the past, he will "punish" the respondent by not initiating a transaction. This is because by behaving dishonestly in the past in a situation virtually identical to the current match, the respondent has revealed his tendency towards dishonest behavior. Unlike in Kandori (1992), however, such punishment behavior is not an outcome of a social norm because, conditional on observing the respondent's past behavior, it is a dominant strategy for the initiator not to initiate (i.e., there is no multiplicity of equilibria and hence an equilibrium selection by a "social norm" to consider).

However, this kind of punishment relies on the perfect observability of individual histories of play (subsequently referred to as "individual histories"). In reality, though, it is often the case that individual histories are observable with noise, or are not observable at all. To be precise, in our model, an individual history is observable with noise if a "spotty history", i.e., a history of dishonest play, generates a "spotty track record" (that is actually observed) with a probability of less than one, and otherwise generates a "clean track record." Consequently, when an initiator is matched with a respondent with a clean track record, he can only rely on the group reputation of the respondents who possess a clean track record. This implies that the imperfect observability of individual histories leads to initiators utilizing both the individual and collective reputations of the respondents when forming beliefs about respondents' trustworthiness. In the extreme case when individual histories are completely unobservable, the respondents' group reputation is the only source of information for the initiators. On the other hand, if individual histories are perfectly observable, the group reputation of the respondents becomes irrelevant.

A general equilibrium is characterized by fractions of initiators who are trusting and by fractions of respondents who are trustworthy, conditional on a particular match situation, which are mutually consistent. We define the level of "trust" as the fraction of initiators who initiate a transaction in a new match when facing a clean track record respondent. We define the level of "trustworthiness" as the fraction of respondents unmatched at the beginning of a typical period who respond honestly to an initiated transaction in a new match. An equilibrium is then essentially an intersection of two behavioral dependencies. The first behavioral dependency characterizes the impact of the degree of trust on the incentive to be trustworthy. An increase in trust increases the expected discounted value associated with being honest regardless of the degree of observability of individual histories. This is because an honest respondent always has a clean track record and has to participate in a new match from time to time. On the other hand, the way an increase in trust affects the expected discounted value associated with being dishonest depends on the degree of the observability of individual histories. If histories are unobservable, every respondent always has a clean track record, and therefore an increase in trust increases the expected discounted value of being dishonest. Because a dishonest respondent is in a new match relatively more frequently than an honest respondent, the increase in the expected discounted value of being dishonest is higher than the corresponding increase for an honest respondent. On the other hand, if individual histories are perfectly observable, the level of trust has no impact on the expected discounted value of being dishonest, since dishonest respondents are never offered an initiated transaction. As a result, an increase in trust makes respondents less likely to behave honestly for low degrees and more likely to behave honestly for high degrees of observability of individual histories of play, while the relationship is non-monotone for intermediate degrees of this observability. However, the result for low degrees of observability crucially depends on a positive probability of the repetition of a successful match. If all the matches last only one period, as in Tirole (1996), and individual histories are completely unobservable, an increase in trust

increases the expected discounted values associated with cheating and behaving honestly in the same way because both types of respondents have to look for new matches in every period and all the respondents always have clean track records. As a result, respondents are no more or less likely to be trustworthy dependent on the level of trust. This distinction illustrates one of the significant new insights originating from the concept of group reputation that one obtains by allowing for repeated matches.

The second behavioral dependency characterizes the impact of the degree of trustworthiness on the incentive to trust. The higher the level of trustworthiness, the more likely initiators are to be matched with honest respondents (even conditionally on observing a clean track record), and therefore the more likely they are to trust.

We integrate these behavioral dependencies, construct a general trust/trustworthiness equilibrium, and prove the existence of the equilibrium. Multiple equilibria involving different levels of trust and trustworthiness in new matches may arise in this economy. In some instances trust and trustworthiness are positively related, which allows for the Pareto ranking of these equilibria, but at other times they may be negatively related, barring any Pareto ranking. However, a generic feature of any equilibrium is that there are mistrusting initiators and untrustworthy respondents who would be better off *ex ante* (before playing the game) if they could commit to trusting and being trustworthy, respectively. We provide some empirical predictions for how the individual pecuniary payoff to trusting depends on the level of trustworthiness, and how the individual pecuniary payoff to being trustworthy depends on the level of trust.

Note that low trust originates from not being able to perfectly observe the individual history of a respondent, and hence the necessity to rely on the group reputation of the respondents. This suggests that the existence of Pareto-dominated equilibria originates from a *negative reputational externality* implicitly embedded in the respondents' group reputation. Essentially, under the missing or imperfect observability of individual histories, the untrustworthy behavior of a particular respondent has a negative impact on the

reputation of the entire group, but this effect is not internalized by the respondent, and the analogous argument applies to trustworthy behavior. As a result, because group reputation is similar in nature to a public good, it is underprovided.

The rest of the paper is structured as follows. Section 2 introduces the model. Section 3 analyzes the partial equilibrium behavior of the initiators when the respondents' trustworthiness is fixed, and analyzes the partial equilibrium behavior of the respondents when the initiators' trust is fixed. Section 4 integrates the two sides of the partial equilibrium analysis into a general equilibrium analysis. Section 5 analyzes the average equilibrium payoffs associated with trusting and mistrusting on the one hand, and being trustworthy and untrustworthy on the other hand. It then proceeds to investigate whether any initiators or respondents would be better off on average if they could commit to an alternative course of action (e.g., trusting rather than not trusting) before playing the game, whether multiple equilibria can be Pareto ranked, and how individual comparative pecuniary payoffs to trust and trustworthiness depend on the average level of trustworthiness and trust, respectively. Section 6 concludes.

2 Model

There is a continuum of individuals with a total measure normalized to 1. The output in the economy is created from business transactions. Each transaction has two parties to it: an initiator and a respondent. Each individual simultaneously participates in both roles. An initiator initiates a transaction by "investing" 1 unit of a generic good, and a respondent, if responding honestly, contributes to the successful completion of the transaction. In such a case the total payoff from the transaction is $2a + 1$ and the net output of $2a$ is shared equally by the two parties, giving a net payoff a to each party. However, the respondent may also respond dishonestly by "stealing" the investment. In such a case the net payoff to the initiator is -1 and the net payoff to the respondent is $1 - d$, where

d is an individual-specific inherent propensity to be honest, measured by the disutility from being dishonest. The value of d has support $[\underline{d}, \overline{d}]$ and with a continuous distribution function F on $[\underline{d}, \overline{d}]$. To make the dishonest response potentially attractive to at least some respondents, we assume that $a < 1 - \underline{d}$. In light of this possibility, an initiator may decide not to initiate a transaction in the first place. In such a case the net payoff to the initiator is $-m$ and the net payoff to the respondent is 0, where m is an individual-specific inherent propensity to trust, captured by the disutility m of mistrust. The value of m has support $[\underline{m}, \overline{m}]$ and a continuous distribution function G on $[\underline{m}, \overline{m}]$. We assume that $\overline{m} < 1$, meaning that none of the initiators are pathological trusters. We also assume that $-\underline{m} < a$, meaning that none of the initiators are pathological mistrusters. In other words, we assume that, in a one-shot game, no initiator would initiate if the respondent is guaranteed to respond dishonestly, and every initiator would initiate if the respondent is guaranteed to respond honestly.[2] If a transaction is not initiated or if an initiated transaction is met with a dishonest response, there is no net output produced (the theft is just a transfer). An extensive form of the transaction game is pictured in Figure 1.[3]

This setup, similar to the investment game analyzed by Berg et al. (1995), tries to capture the notions of trust and trustworthiness. Successful completion of a transaction requires both the trusting approach of the initiator and the trustworthy approach of the respondent. If either is missing, the transaction fails and no net output is produced (although some existing wealth might be redistributed).

Each period a subgroup of initiators interacts with a subgroup of respondents by participating in an initiator-respondent match. Even though each individual has a dual role

[2]Note that we do not assume that d and m are distributed independently across individuals. Indeed, they may be correlated. Whether they are correlated or not, however, is immaterial to the subsequent analysis since each individual acts independently in his initiator and respondent roles.

[3]Note that the fact that the participants in the transaction share the net gain equally when the transaction is successfully completed and that the initiator's pecuniary loss is equal to the respondent's pecuniary gain under a dishonest response is inconsequential (as long as the net gain is shared in fixed proportions), because the behavior of each group is only affected by their own payoff structure. In other words, all that matters for a particular initiator are the magnitudes of a and m relative to the amount of the investment necessary to initiate the transaction. Similarly, all that matters for a particular respondent are the magnitudes of a and d relative to the amount that can be stolen when responding dishonestly.

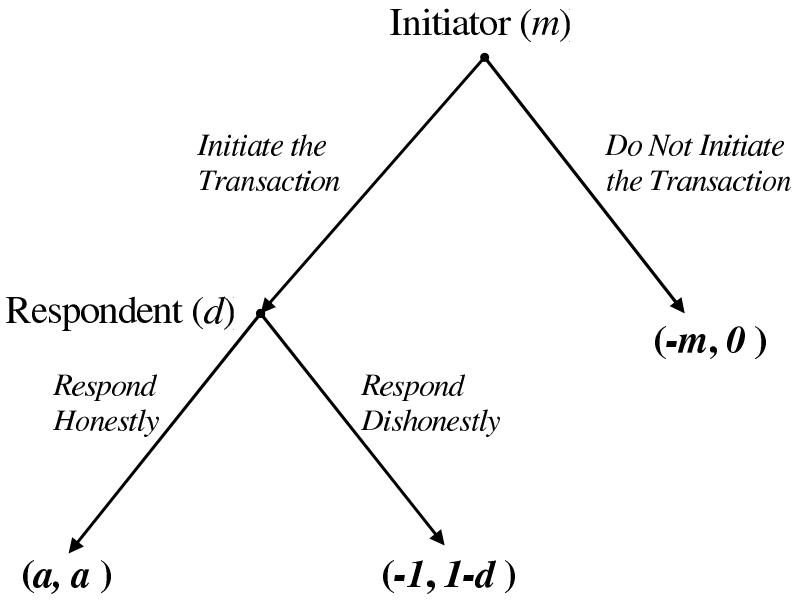

FIGURE 1: Extensive form of the transaction game

in each period, acting both as an initiator and a respondent, it is helpful to separate these two roles and to think of the initiators and the respondents as two separate groups of the same measure.[4] At the beginning of each period there are equally sized groups of matched initiators and matched respondents and equally sized groups of unmatched initiators and unmatched respondents. Those matched participate in their "surviving" matches from the previous period. Each unmatched initiator gets matched with probability $\beta \in (0, 1]$ to some unmatched respondent and vice versa. Then, by the law of large numbers, β is also the fraction of both the searching initiators and the searching respondents who get matched in a new match in the current period. If an initiator or a respondent is unmatched, his or her payoff for the current period is 0. If an initiator and a respondent are matched (in a new or a surviving match), they play the stage game outlined above and collect their payoffs. If the transaction is completed successfully (i.e., it is initiated and responded to honestly), the match survives to the next period with probability $\alpha \in [0, 1)$. Otherwise it is dissolved and both participants enter the next period unmatched. This is also the case if the transaction is completed successfully but the match does not survive until the next

[4]This simplification is innocuous since meeting oneself is a zero probability event.

period for exogenous reasons, which happens with the conditional probability $1-\alpha$. In turn, α is then also the fraction of matches with successfully completed transactions that survive to the next period. Intuitively, even if the match is "working", exogenous events such as population mobility or business turnover may cause the match to break up. All individuals are risk neutral and have a discount factor $\delta \in [0,1)$.

In addition to destroying the current match, dishonest behavior also has consequences for the individual reputation of the respondent. In particular, if the respondent has cheated in the most recent initiated transaction, which we refer to as having a "spotty history", this fact gets revealed to the initiator in the respondent's next match with probability $x \in [0,1]$, giving this respondent a "spotty track record". However, with probability $1-x$, the respondent's spotty history does not get revealed, in which case the initiator is observationally indistinguishable from a respondent who has a "clean history", i.e., a respondent who has responded honestly in his most recent transaction. If a respondent has a clean history or a spotty history that is not revealed, the respondent acquires a "clean track record". When $x=1$, individual histories are perfectly observable. At the other extreme, when $x=0$, individual histories are unobservable. In the intermediate case when $x \in (0,1)$, individual histories are observable with noise.

To make the analysis tractable, we restrict our attention to steady states and make the following additional assumptions:

Assumption 1. *Initiators and respondents condition their strategies on the age of the match only to the extent that it is new or surviving.*[5]

[5]This assumption is made in order to simplify the analysis. Given that a necessary condition for match survival is the successful completion of the transaction in each period, a general strategy space would allow strategies to condition on the age of the match, since that is the only variable that may differ from one surviving match to another. Indeed, one could envision an equilibrium in which, conditional on match survival, initiators of type m initiate until period $x(m)$ and respondents of type d respond honestly until period $y(d)$, where $x(\cdot)$ and $y(\cdot)$ are (weakly) increasing and potentially infinitely valued. In such an equilibrium, given the age of a particular match, optimal initiator and respondent decisions would be determined by the intrinsic behavioral propensities and the updated distributions of match partner types (where the support of the latter only includes opponent types whose strategies prescribe cooperation until at least the realized age of the match). Intuitively, we focus on equilibria where $x(\cdot)$ and $y(\cdot)$ only assume values of 0, 1, or infinity. We believe that the subclass of strategies we focus on sufficiently captures the essentials of trust

Assumption 2. *If an initiator is indifferent between initiating and not initiating, she chooses to initiate. If a respondent is indifferent between responding honestly and dishonestly, he chooses to respond honestly.*

In the next section, we fix the behavior of the respondents and derive the induced behavior of the initiators, and then fix the behavior of the initiators and derive the induced behavior of the respondents. Section 4 then aggregates these individual decisions to determine a general trust/trustworthiness equilibrium.

3 Partial Equilibrium Analysis

In each period, after an initiator has realized whether she is matched, whether the match (if any) is new or surviving, after observing the track record of her opponent (if the match is new), and given a vector θ_R of summary statistics describing the behavior of the respondents, an initiator of type m may find herself in one of the following four states with the associated expected discounted payoff values:

I1: Not matched: $I_1(m, \theta_R)$

I2: Matched in a new match with a respondent having a spotty track record: $I_2(m, \theta_R)$

I3: Matched in a new match with a respondent having a clean track record: $I_3(m, \theta_R)$

I4: Matched in the second period of a surviving match: $I_4(m, \theta_R)$

I5: Matched in the third or higher period of a surviving match: $I_5(m, \theta_R)$.

Based on the realization of whether he has cheated in his most recent initiated transaction, whether he is matched, whether the match is new or surviving, whether the initiator has initiated a transaction, whether his track record is clean or spotty, and given a vector θ_I of summary statistics describing the behavior of the initiators, a respondent of type d may find himself in one of the following states with the associated expected discounted payoff values:

and trustworthiness in an equilibrium setting.

R1: Not matched or matched without an initiated transaction, having cheated in the most recent initiated transaction: $R_1(d, \theta_I)$

R2: Not matched or matched without an initiated transaction, having not cheated in the most recent initiated transaction: $R_2(d, \theta_I)$

R3: Matched in a new match with an initiated transaction and a spotty track record: $R_3(d, \theta_I)$

R4: Matched in a new match with an initiated transaction and a clean track record: $R_4(d, \theta_I)$

R5: Matched in a surviving match with an initiated transaction: $R_5(d, \theta_I)$

The summary statistics vector θ_R describes the average respondents' behavior. In particular, let $k_3, k_4 \in [0, 1]$ be the probabilities that a randomly chosen respondent, conditional on this respondent achieving states R3 and R4, respectively, will behave honestly in that state. Let $k_5 \in [0, 1]$ be the probability that a randomly chosen respondent, conditional on this respondent achieving state R5, will behave honestly in that state *in the second period of the given match*.[6] Also let $q \in [0, 1]$ be the probability that a randomly chosen respondent who is unmatched at the beginning of a typical period has a clean history. Then $\theta_R = (k_3, k_4, k_5, q)$.

Similarly, θ_I describes the average initiator behavior. In particular, let $s_2, s_3, s_4 \in [0, 1]$ be the probabilities that a randomly chosen initiator, conditional on this initiator achieving states I2, I3, and I4, respectively, will initiate a transaction in that state. If in state I5, Assumption 1 implies that an initiator's action is perfectly revealed by her behavior in the previous period, i.e., she initiates a transaction. Then $\theta_I = (s_2, s_3, s_4)$.

Given this setup, first consider the decision making of an initiator m given θ_R (for simplicity of notation we omit θ_R from the list of arguments in the value functions). In state I1, there is no current decision to make and the initiator collects a payoff of 0 in the current period. In the next period she remains unmatched (i.e., remains in state I1) with

[6] If the match lasts three or more periods, Assumption 1 implies that a respondent's action is perfectly revealed by his behavior in the previous period, i.e., he responds honestly.

probability $1-\beta$, and gets matched in a new match with probability β. In the latter case, the probability of being matched with a respondent with a spotty track record (i.e., getting to state I2) is $x(1-q)$, while the probability of being matched with a respondent with a clean track record (i.e., getting to state I3) is $q+(1-x)(1-q)$, or $1-x+qx$. Therefore the Bellman equation for $I_1(m)$ is

$$I_1(m) = 0 + \delta\left\{(1-\beta)I_1(m) + \beta\left[x(1-q)I_2(m) + (1-x+qx)I_3(m)\right]\right\}. \quad (1)$$

In state I2, not initiating leads to a current payoff of $-m$ and a continuation value that is equivalent to being in state I1 currently. On the other hand, if initiating, the respondent replies honestly with probability k_3, leading to a current payoff of a. In addition, the match then survives with probability α, putting the initiator into state I4 in the following period, and it does not survive with probability $1-\alpha$, leading to a continuation payoff that is equivalent to being in state I1 currently. The respondent replies dishonestly with probability $1-k_3$, leading to a current payoff of -1 and a continuation payoff that is equivalent to being in state I1 currently. Therefore the Bellman equation for $I_2(m)$ is

$$\begin{aligned} I_2(m) = \max\{&-m + I_1(m); k_3\left[a + \alpha\delta I_4(m) + (1-\alpha)I_1(m)\right] \\ &+(1-k_3)\left[-1 + I_1(m)\right]\}. \end{aligned} \quad (2)$$

The first term in the maximand corresponds to the value of not initiating, while the second term corresponds to the value of initiating.

Analogous reasoning for states I3 and I4 leads to Bellman equations for $I_3(m)$ and $I_4(m)$ of the form

$$\begin{aligned} I_3(m) = \max\{&-m + I_1(m); k_4\left[a + \alpha\delta I_4(m) + (1-\alpha)I_1(m)\right] \\ &+(1-k_4)\left[-1 + I_1(m)\right]\}, \end{aligned} \quad (3)$$

and

$$I_4(m) = \max\{-m + I_1(m); k_5\left[a + \alpha\delta I_5(m) + (1-\alpha)I_1(m)\right] + (1-k_5)\left[-1 + I_1(m)\right]\}. \quad (4)$$

Again, in both (3) and (4), the first term in the maximand corresponds to the value of not initiating, while the second term corresponds to the value of initiating.

In state I5, Assumption 1 implies that the initiator's strategy is to initiate a transaction, which will be followed by an honest response from the respondent. Therefore the initiator collects a current payoff of a. In addition, the match survives with probability α, putting the initiator into state I5 in the following period, and it does not survive with probability $1-\alpha$, leading to a continuation payoff that is equivalent to being is state I1 currently. Therefore the Bellman equation for $I_5(m)$ is

$$I_5(m) = a + \alpha\delta I_5(m) + (1-\alpha)I_1(m). \quad (5)$$

Now consider the decision making of a respondent d given θ_I (for simplicity of notation, we omit θ_I from the list of arguments in the value functions). In state R1, there is no current decision to make and the respondent collects a payoff of 0 in the current period. In the next period he remains unmatched (i.e., remains in state R1) with probability $1-\beta$, and gets matched in a new match with probability β. In the latter case, his spotty history is revealed with probability x. In this case the initiator, now in state I2, will initiate with probability s_2, putting the respondent into state R3 in the next period, and does not initiate with probability $1-s_2$, putting the respondent back into state R1 in the next period. However, with probability $1-x$, the respondent's spotty history is not revealed. In this case the initiator, now in state I3, will initiate with probability s_3, putting the respondent into state R4 in the next period, and does not initiate with probability $1-s_3$, putting the respondent

back into state R1 in the next period. Therefore the Bellman equation for $R_1(d)$ is

$$\begin{aligned} R_1(d) = 0 + \delta \{ & [1 - \beta + \beta (x(1 - s_2) + (1 - x)(1 - s_3))] R_1(d) \\ & + \beta x s_2 R_3(d) + \beta (1 - x) s_3 R_4(d) \} . \end{aligned} \tag{6}$$

In state R2, there is no current decision to make and the respondent collects a payoff of 0 in the current period. In the next period, he remains unmatched (i.e., in state R2) with probability $1 - \beta$, and gets matched with probability β. In the latter case, because the respondent's track record is necessarily clear, the initiator, now in state I3, will initiate with probability s_3, putting the respondent into state R4 in the next period, and does not initiate with probability $1-s_3$, putting the respondent back into state R2 in the next period. Therefore the Bellman equation for $R_2(d)$ is

$$R_2(d) = 0 + \delta \{ (1 - \beta s_3) R_2(d) + \beta s_3 R_4(d) \} . \tag{7}$$

In state R3, responding dishonestly leads to a current payoff of $1-d$ and a continuation value that is equivalent to being in state R1 currently. On the other hand, responding honestly leads to a current payoff of a. The match then survives until the next period with probability α. In that period, the initiator will initiate with probability s_4, putting the respondent into state R5, and she will not initiate with probability $1 - s_4$, putting the respondent into state R2. With probability $1 - \alpha$, the match will not survive until the following period, leading to a continuation payoff that is equivalent to being in state R2 currently. Therefore the Bellman equation for $R_3(d)$ is

$$\begin{aligned} R_3(d) = \max \{ & 1 - d + R_1(d); a + \alpha \delta [s_4 R_5(d) + (1 - s_4) R_2(d)] \\ & + (1 - \alpha) R_2(d) \} . \end{aligned} \tag{8}$$

The first term in the maximand corresponds to the value of responding dishonestly, while

the second term corresponds to the value of responding honestly.

Analogous reasoning for state R4 implies that

$$\begin{aligned} R_4(d) &= \max\left\{1-d+R_1(d); a+\alpha\delta\left[s_4R_5(d)+(1-s_4)R_2(d)\right]+(1-\alpha)R_2(d)\right\} \\ &= R_3(d). \end{aligned} \tag{9}$$

This result implies that any given respondent would behave identically in states R3 and R4. However, because the composition of the respondent groups with respect to their predisposition to be trustworthy is in general different across these two states, this result does not imply that $k_3 = k_4$.

Similar reasoning, with one modification, also applies to state R5. The modification stems from the fact that once in state R5, Assumption 1 implies that the initiator's strategy is to initiate a transaction in the following period if the match survives until then. As a result, s_4 in equation (8) is replaced by 1. Therefore the Bellman equation for $R_5(d)$ is

$$R_5(d) = \max\left\{1-d+R_1(d); a+\alpha\delta R_5(d)+(1-\alpha)R_2(d)\right\}. \tag{10}$$

Again, the first term in the maximand corresponds to the value of responding dishonestly, while the second term corresponds to the value of responding honestly.

Equations (1) to (5) completely characterize the behavior of initiators given θ_R, and equations (6) to (10) completely characterize the behavior of respondents given θ_I. These are the partial equilibrium characterizations. In the next section we combine the behavior of the initiators with the behavior of the respondents to derive a general equilibrium.

4 General Equilibrium Analysis

4.1 Definition and Characterization

We begin by defining what we mean by a general trust/trustworthiness equilibrium.

Definition 1. *A **general trust/trustworthiness equilibrium** is a mutually consistent combination of $\theta_I = (s_2, s_3, s_4)$ and $\theta_R = (k_3, k_4, k_5, q)$ for which individual initiator and respondent behavior is driven by the choices implicit in (1)-(5) and (6)-(10).*

Combining equations (7), (9), and (10) gives the following result (see the Appendix for a proof):

Lemma 1. *All respondents, conditional on achieving state R5, respond honestly in this state. That is, $k_5 = 1$.*

Intuitively, if in state R5, the respondent must have chosen to respond honestly in the first period of the given match, i.e., in state R3 or R4, even in the presence of uncertainty about whether the initiator would or would not initiate in the following period when in state I4. It then follows that the respondent will also opt to respond honestly once it is certain that the initiator will initiate in the next period (and every subsequent period as long as the match lasts).

If a respondent has a clean history, this history might originate from any of the states R3, R4, or R5. However, because being in state R5 is always preceded by responding honestly in state R3 or state R4, and because (9) implies that any given respondent would behave identically in these two states, the strategy of a clean history respondent must prescribe an honest response in states R3 and R4. On the other hand, Lemma 1 implies that respondents never acquire a spotty history by acting dishonestly in state R5. Therefore a respondent can only acquire a spotty history by acting dishonestly in state R3 or state R4. Equation (9), however, implies that any given respondent would behave identically in states R3 and R4. Therefore, regardless of whether a respondent acquired a spotty history

in state R3 or R4, his strategy must prescribe a dishonest response in both of these states. This leads to the following lemma:

Lemma 2. *A respondent responds honestly in state R4 if and only if he has a clean history.*

In addition, because anyone with a spotty record has a spotty history, any respondent who achieves state R3 is going to respond dishonestly in that state. This implies the following result:

Lemma 3. *All respondents, conditional on achieving state R3, respond dishonestly in this state. That is,* $k_3 = 0$.

Lemma 3 implies that (2) is reduced to

$$I_2(m) = \max\left\{-m + I_1(m); -1 + I_1(m)\right\}.$$

That is, when in state I2, an initiator faces a choice between not trusting and being cheated. Because we assumed that nobody is a pathological truster ($\overline{m} < 1$), it is better to not trust, which also implies that

$$I_2(m) = -m + I_1(m). \tag{11}$$

This result is summarized by the following lemma:

Lemma 4. *All initiators, conditional on achieving state I2, do not trust in this state. That is,* $s_2 = 0$.

Combining (3)-(5) with Lemmata 1 and 3 gives the following result (see the Appendix for a proof):

Lemma 5. *All initiators, conditional on achieving state I4, trust in this state. That is,* $s_4 = 1$.

The intuition behind this result is similar to the intuition underlying Lemma 1. If he is in state I4, the initiator must have chosen to initiate in the first period of the given match,

i.e., in state I3,[7] even in the presence of uncertainty about whether the respondent would respond honestly. It then follows that the initiator will also opt to initiate once it is certain that the respondent will respond honestly (and will keep doing so as long as the match lasts).

Lemma 2 implies that k_4 is equal to the probability that a randomly chosen respondent who has achieved state R4 has a clean history. Because every unmatched respondent has an equal chance of being matched with an initiator, the probability that a randomly chosen newly matched respondent has a clean history is equal to q. From the newly matched respondents, the ones who have a spotty history acquire a spotty track record with probability x. Therefore the probability that a newly matched respondent has a clean track record is $q + (1-x)(1-q) = 1 - x + qx$. As a result, the probability that a newly matched respondent has a clean history conditional on him having a clean track record is $q/(1-x+qx)$. Because every newly matched respondent with a clean track record is equally likely to be matched with a trusting initiator, i.e., to reach state R4, Lemma 2 implies that

$$k_4 = \frac{q}{1-x+qx}. \tag{12}$$

This computation fails in the pathological case when $x = 1$ *and* $q = 0$ because the conditioning set (newly matched respondents with clean track records) has measure zero. In such a case no respondent ever achieves state R4 because all the newly matched respondents have spotty histories, which get perfectly revealed in their track records. To rule out this pathological case, we impose the following assumption:

Assumption 3. *If $x = 1$, F assigns a positive measure to the set $\left[1 - \frac{a}{1-\alpha\delta}, \infty\right)$. In addition, every respondent is assumed to possess a clean history in equilibrium if he would respond honestly in state R4.*

The first part of this assumption implies that there will always be a positive measure

[7] Note that, by Lemma 4, no initiators initiate a transaction in state I2.

of respondents who would respond honestly in state R4 *if in that state*.[8] The second part of this assumption rules out the case when even though there are respondents who would respond honestly in state R4 if in that state, they have spotty track records because of an initial assignment of such track records, and therefore never get a chance to "clean" themselves by responding honestly since $s_2 = 0$. The two parts of Assumption 3 then imply that $q > 0$ when $x = 1$.[9]

Now consider the behavior of an initiator in state I3. Using (1), (3), (4), (5), (11), and Lemmata 1 and 5, we obtain the following result (see the Appendix for a proof):

Lemma 6. *An initiator m initiates a transaction in state I3 if and only if $m \geq m(q, \delta)$, where $m : [0, 1]^2 \to R$ is defined by*

$$m(q, \delta) \equiv \frac{(1 - \alpha\delta)(1 - x)(1 - q) - aq}{(1 - \alpha\delta)(1 - x + qx) + \alpha\beta\delta q}.$$

Lemma 6 says that, in a new match with a respondent with a clean record, initiators with a relatively high level of intrinsic trust will behave in a trusting way and initiate, while initiators with a relatively low level of intrinsic trust will not trust and thus will not initiate. For future reference, we label the former as "trusting initiators," and the latter as

[8]Consider a respondent d in state R4. Responding dishonestly generates a current payoff of $1 - d$ and a spotty history. However, because $x = 1$, the spotty history leads to a spotty track record that cannot be cleaned in the future since $s_2 = 0$ (see Lemma 4). So the expected discounted payoff of responding dishonestly is $1 - d$. The expected discounted payoff of responding honestly consists of the expected discounted payoff from the current match and an expected continuation value from engaging in new matches in the future. The latter is necessarily nonnegative since it is always possible to avoid negative payoffs by responding honestly. Discounting for time and the survival probability of the current match, the expected discounted payoff from the current match is equal to $a/(1 - \alpha\delta)$. This is then also a lower bound for the expected discounted payoff to responding honestly. Consequently, if

$$\frac{a}{1 - \alpha\delta} \geq 1 - d \;\Leftrightarrow\; d \geq 1 - \frac{a}{1 - \alpha\delta},$$

then the respondent will respond honestly.

[9]The two parts of Assumption 3 imply that the set of matched and unmatched respondents with a clean history has a positive measure. This implies that the set of unmatched respondents with a clean history must have a positive measure, and hence $q > 0$. Suppose this was not the case. Then the set of matched respondents with a clean history must have a positive measure. But since the fraction $1 - \alpha$ of matches involving these respondents do not survive until the following period, there must be a positive measure of unmatched respondents with a clean history in the following period. Because we are focusing on a steady state, this is a contradiction.

"mistrusting initiators." The threshold $m(q, \delta)$ is a strictly decreasing function of q, the probability that a randomly chosen unmatched initiator has a clean history. Intuitively, the more likely the unmatched respondents are to have a clean history, the higher the probability that a randomly chosen initiator would trust if in state R4 because the clean track record respondents are less likely to include "false negatives," i.e., respondents with a spotty history but a clean track record.

Finally, consider the behavior of a respondent in state R4. Using (6), (9), (10), and Lemmata 1, 4, and 5, we obtain the following result (see the Appendix for a proof):

Lemma 7. *A respondent d responds honestly in state R4 if and only if $d \geq d(s_3, \delta)$, where $d : [0,1]^2 \to R$ is defined by*

$$d(s_3, \delta) \equiv 1 - \frac{1 - \delta + \beta\delta s_3}{(1 - \alpha\delta + \alpha\beta\delta s_3)\left[1 - \delta + \beta\delta s_3(1 - x)\right]} a.$$

Lemma 7 says that, in a new match with a clean track record, respondents with a relatively high level of intrinsic honesty will behave in a trustworthy way and reply honestly, while respondents with relatively low intrinsic honesty will not behave in a trustworthy way and will reply dishonestly. For future reference, we label the former as "trustworthy respondents," and the latter as "untrustworthy respondents." This behavioral pattern is due to the fact that the latter group will find theft attractive because of their low "moral barriers", even though it entails termination of the match and potential damage to their individual reputation. On the other hand, the former group will not find theft attractive because of their high "moral barriers" and their individual reputation considerations. The threshold $d(s_3, \delta)$ is, however, in general nonmonotone in s_3. If $x = 0$ (i.e., the individual histories are unobservable and only the group reputation of respondents matters in state I3), the threshold is strictly increasing in s_3. This means that the more trusting the initiators are in state I3, the *lower* the probability that a randomly chosen respondent would respond honestly if in state R4. However, this effect is present only if $\alpha > 0$, i.e., only

if successful matches are repeated at least some of the time. When $\alpha = 0$, a dishonest behavior has no cost in terms of destroying the current match. Its only cost stems from potentially damaging one's reputation, a consideration which is not present when $x = 0$. Compared to Tirole (1996), who only allows for one-period matches, this finding represents a significant new insight originating from the concept of group reputation that one obtains by allowing for repeated matches. On the other hand, if $x = 1$ (i.e., individual histories are perfectly observable), the threshold is strictly decreasing in s_3. This means that the more trusting the initiators are in state I3, the *higher* the probability that a randomly chosen respondent would respond honestly if in state R4. This dichotomy extends the discussion in the introduction about the way trust impacts respondents' decisions to be trustworthy under various degrees of observability of individual histories.

In order to complete the characterization of general equilibrium, we need to link the individual behavior captured by Lemmata 6 and 7 with s_3 and q, respectively. Let h^I be the fraction of trusting initiators that are unmatched at the beginning of a typical period. Because every unmatched initiator has an equal chance of being matched, and every matched initiator is equally likely to be matched with a clean track record respondent, the probability s_3 that a randomly chosen initiator who has achieved state I3 is trusting is equal to the probability that a randomly chosen unmatched initiator is trusting. By Lemma 6, the measure of trusting initiators is $1 - G[m(q, \delta)]$, and hence the measure of unmatched trusting initiators is $h^I \{1 - G[m(q, \delta)]\}$. On the other hand, the measure of mistrusting initiators is $G[m(q, \delta)]$, and all of these mistrusting initiators are unmatched in every period since they never participate in a surviving match. Therefore

$$s_3 = \frac{h^I \{1 - G[m(q, \delta)]\}}{G[m(q, \delta)] + h^I \{1 - G[m(q, \delta)]\}}. \tag{13}$$

Similarly, let h^R be the fraction of trustworthy respondents that are unmatched at the beginning of a typical period in a steady state. Because all trustworthy respondents have

clean histories and all untrustworthy respondents have spotty histories, q, the probability that a randomly chosen unmatched respondent has a clean history is equal to the fraction of unmatched trusting respondents among all the unmatched respondents. By Lemma 7, the measure of trustworthy respondents is $1 - F[d(s_3, \delta)]$, and hence the measure of unmatched trustworthy respondents is $h^R\{1 - F[d(s_3, \delta)]\}$. On the other hand, the measure of untrustworthy respondents is $F[d(s_3, \delta)]$, and all of these untrustworthy respondents are unmatched in every period since they never participate in a surviving match. Therefore

$$q = \frac{h^R\{1 - F[d(s_3, \delta)]\}}{F[d(s_3, \delta)] + h^R\{1 - F[d(s_3, \delta)]\}}. \tag{14}$$

Note that s_3 falls short of $1 - G[m(q, \delta)]$, which is the probability that a randomly drawn initiator (not necessarily in state I3) is trusting. Intuitively, this is because trusting initiators are less likely to find themselves in state I3 relative to mistrusting initiators, since the former are more likely to participate in surviving matches. Similarly, q falls short of $1 - F[d(s_3, \delta)]$, which is the probability that a randomly drawn respondent (not necessarily unmatched) is trustworthy, and hence has a clean history. Intuitively, this is because trustworthy respondents are less likely to find themselves unmatched relative to untrustworthy respondents, since the former are more likely to participate in surviving matches.

In the final step of deriving a general equilibrium, we need to find h^I and h^R. Since h^I has to stay constant over time, in any period the measure of new matches involving trusting initiators that survive until the following period has to be equal to the measure of surviving matches involving trusting initiators that get dissolved in the current period. As for the former, the fraction h^I of trusting initiators (those unmatched) results in the fraction βh^I of trusting initiators involved in new matches, the fraction $(1 - x + qx)\beta h^I$ of trusting initiators involved in new matches experiencing an initiated transaction (all those who achieve state I3), the fraction $k_4(1 - x + qx)\beta h^I$ of trusting initiators involved in

new matches experiencing a successfully completed transaction, and finally, the fraction $\alpha k_4 (1 - x + qx) \beta h^I$ of trusting initiators involved in new matches that survive until the following period. Using (12), this fraction is equal to $\alpha\beta q h^I$. As for the latter, the fraction $1 - h^I$ of trusting initiators (those participating in surviving matches) results in the fraction $(1 - \alpha)(1 - h^I)$ of trusting initiators whose matches get dissolved in the current period. In equilibrium, then, $\alpha\beta q h^I = (1 - \alpha)(1 - h^I)$, which gives

$$h^I = \frac{1 - \alpha}{1 - \alpha + \alpha\beta q}. \tag{15}$$

Using this result for substitution into (13) then gives

$$s_3 = \frac{(1 - \alpha) \{1 - G [m(q, \delta)]\}}{1 - \alpha + \alpha\beta q G [m(q, \delta)]}. \tag{16}$$

Similarly, since h^R has to stay constant over time in a steady state, in any period the measure of new matches involving trustworthy respondents that survive until the following period has to be equal to the measure of surviving matches involving trustworthy respondents that get dissolved in the current period. As for the former, the fraction h^R of trustworthy respondents (those unmatched) results in the fraction βh^R of trustworthy respondents involved in new matches, the fraction $s_3 \beta h^R$ of trustworthy respondents involved in new matches experiencing an initiated transaction (because all of the trustworthy respondents have a clean history and hence also a clean track record, they all achieve state I3 once in a new match), the fraction $s_3 \beta h^R$ of trustworthy respondents involved in new matches experiencing a successfully completed transaction and finally, the fraction $\alpha s_3 \beta h^R$ of trustworthy respondents involved in new matches that survive until the following period. As for the latter, the fraction $1 - h^R$ of trustworthy respondents (those participating in surviving matches) results in the fraction $(1 - \alpha)(1 - h^R)$ of trustworthy respondents participating in surviving matches that get dissolved in the current period. In

equilibrium, then, $\alpha s_3 \beta h^R = (1-\alpha)(1-h^R)$, which gives

$$h^R = \frac{1-\alpha}{1-\alpha+\alpha\beta s_3}. \tag{17}$$

Using this result for substitution into (14) then gives

$$q = \frac{(1-\alpha)\left\{1-F\left[d(s_3,\delta)\right]\right\}}{1-\alpha+\alpha\beta s_3 F\left[d(s_3,\delta)\right]}. \tag{18}$$

Lemmata 1 to 7 together with (12), (16), and (18) completely characterize any general trust/trustworthiness equilibrium. These results are summarized in the following proposition:

Proposition 1. *A general trust/trustworthiness equilibrium satisfies* $s_2 = 0$, $s_4 = 1$, $k_3 = 0$, $k_5 = 1$,

$$s_3 = s(q) \equiv \frac{(1-\alpha)\left\{1-G\left[m(q,\delta)\right]\right\}}{1-\alpha+\alpha\beta q G\left[m(q,\delta)\right]},$$

$$q = q(s_3) \equiv \frac{(1-\alpha)\left\{1-F\left[d(s_3,\delta)\right]\right\}}{1-\alpha+\alpha\beta s_3 F\left[d(s_3,\delta)\right]},$$

and

$$k_4 = \frac{q}{1-x+qx}.$$[10]

[10] At this point the reader may wonder whether the measure of unmatched initiators is equal to the measure of unmatched respondents, i.e., whether

$$G\left[m(q,\delta)\right] + h^I\left\{1-G\left[m(q,\delta)\right]\right\} = F\left[d(s_3,\delta)\right] + h^R\left\{1-F\left[d(s_3,\delta)\right]\right\}. \tag{19}$$

The answer is yes. To see why, note that (16) implies that

$$G\left[m(q,\delta)\right] = \frac{(1-\alpha)(1-s_3)}{1-\alpha+\alpha\beta s_3 q}, \tag{20}$$

and (18) implies that

$$F\left[d(s_3,\delta)\right] = \frac{(1-\alpha)(1-q)}{1-\alpha+\alpha\beta s_3 q}. \tag{21}$$

Using these results together with (15) and (17), both sides of (19) are equal to

$$h = \frac{1-\alpha}{1-\alpha+\alpha\beta s_3 q}, \tag{22}$$

where h denotes a common equilibrium measure of unmatched initiators and respondents.

In the remainder of the paper we identify the equilibrium level of "trust" with s_3 and we identify the equilibrium level of "trustworthiness" with q.[11]

4.2 Existence and Multiplicity of Equilibria

Having characterized a general trust/trustworthiness equilibrium, a natural question is whether such an equilibrium exists. A simple fixed point argument shows that this is indeed the case (see the Appendix for a proof).

Proposition 2. *There exists a general trust/trustworthiness equilibrium.*

Although a general trust/trustworthiness equilibrium always exists, it may not be unique. This non-uniqueness result parallels the results of Rosenthal (1979) and Tirole (1996). The possibility of multiple equilibria is illustrated by Examples 1 and 2 presented in the Appendix. Example 1 illustrates a case in which trust and trustworthiness are positively related. On the other hand, Example 2 illustrates a case in which trust and trustworthiness are negatively related. The comparison of the two examples invites the question of whether various multiple equilibria can be Pareto ranked. The next section shows that this is not the case in general, precisely because of situations like the one illustrated in Example 2. However, if trust and trustworthiness are positively related as in Example 1, various equilibria can be Pareto ranked, with equilibria with a higher level of trust and trustworthiness being Pareto superior.

5 Extensions

In this section, we use our model to investigate several questions. First, would any initiators or respondents gain in equilibrium if they could *ex ante*, i.e., before knowing which

[11] Trustworthiness could alternatively be measured by k_4. However, since (12) implies that there is a one-to-one relationship between k_4 and q (except when $x = 1$, in which case $k_4 = 1$ regardless of the value of q) and it is more convenient to carry out the analysis in terms of q, we use the latter as our measure of trustworthiness.

state they are in, commit to an alternative course of action in state I3 or state R4, respectively? Second, are various multiple equilibria Pareto ranked? Third, how does the comparative pecuniary payoff to trust (excluding the disutility from mistrust) depend on the trustworthiness of the respondents, and how does the comparative pecuniary payoff to trustworthiness (excluding the disutility from being untrustworthy) depend on the initiators' trust? We are especially interested in the last question because it provides a prediction that can be empirically tested, as in Slemrod and Katuscak (2004). In order to address these questions, we first develop two measures of welfare based on expected per period total utility payoff and expected per period pecuniary payoff.

5.1 Individual Payoffs

Let s_3^* and q^* be the equilibrium levels of trust and trustworthiness, respectively. We focus on two measures of individual payoffs: the expected per period total utility payoff, denoted $\widehat{\Pi}$, and the expected per period pecuniary payoff, denoted Π. Both of these expected payoffs depend on the range of outcome situations that an initiator or respondent may be in at the end of a typical period, as well as the equilibrium probability distribution over these outcomes. Note that the concept of outcome is different from the concept of state. While states are various *ex ante* decision making situations, outcomes are various *ex post* payoff situations.

At the end of a period, an initiator or a respondent may be in four different outcome situations at the end of a typical period. First, she may be unmatched. Second, she may be matched in a match without an initiated transaction. Third, she may be matched with an initiated transaction, but with a dishonest response. Fourth, she may be matched in a match with a successfully completed transaction. Let the steady state probabilities of these four outcomes be, in the same order, $\pi^I(m, q^*, i)$ and $\pi^R(d, s_3^*, i)$, $i \in \{1, 2, 3, 4\}$, for initiators of type m and respondents of type d, respectively. For an initiator, the per period total utility payoffs are 0, $-m$, -1, and a, and the per period pecuniary payoffs

are 0, 0, -1, and a in the four respective outcomes. For a respondent, the per period total utility payoffs are 0, 0, $1-d$, and a, and the per period pecuniary payoffs are 0, 0, 1, and a in the four respective outcomes. Hence the expected per period total utility and pecuniary payoffs for an initiator of type m, denoted $\widehat{\Pi}^I(m,q^*)$ and $\Pi^I(m,q^*)$, respectively, are given by

$$\widehat{\Pi}^I(m,q^*) = -\pi^I(m,q^*,2)m - \pi^I(m,q^*,3) + \pi^I(m,q^*,4)a, \tag{23}$$

and

$$\Pi^I(m,q^*) = -\pi^I(m,q^*,3) + \pi^I(m,q^*,4)a. \tag{24}$$

Similarly, the expected per period total utility and pecuniary payoffs for a respondent of type d, denoted $\widehat{\Pi}^R(d,s_3^*)$ and $\Pi^R(d,s_3^*)$, respectively, are given by

$$\widehat{\Pi}^R(d,s_3^*) = \pi^R(d,s_3^*,3)(1-d) + \pi^R(d,s_3^*,4)a, \tag{25}$$

and

$$\Pi^R(d,s_3^*) = \pi^R(d,s_3^*,3) + \pi^R(d,s_3^*,4)a. \tag{26}$$

In order to compute these expected payoffs, we need to find the steady state equilibrium probability distribution over the four outcomes for each initiator and for each respondent. These probability distributions will differ across various initiators only to the extent of whether they trust or do not trust, and they will differ across various respondents only to the extent of whether they are trustworthy or untrustworthy. Among the trusting initiators, the measure h^I are unmatched at the beginning of a typical period. Of these, βh^I become matched in a new match, while $(1-\beta)h^I$ remain unmatched throughout the period. Therefore $\pi^I(\text{trust},q^*,1) = (1-\beta)h^I$. Of the βh^I trusting initiators matched in a new match, $[q^* + (1-x)(1-q^*)]\,\beta h^I = (1-x+q^*x)\,\beta h^I$ face a respondent with a clean track record (state I3), and hence initiate a transaction,

while the remaining $x(1-q^*)\beta h^I$ face a respondent with a spotty track record and hence do not initiate a transaction. Therefore $\pi^I(\text{trust},q^*,2) = x(1-q^*)\beta h^I$. Of the $(1-x+q^*x)\,\beta h^I$ trusting initiators matched in a new match with an initiated transaction, $(1-k_4^*)\,(1-x+q^*x)\,\beta h^I = (1-x)(1-q^*)\beta h^I$ (using (12)) experience a dishonest response, yielding $\pi^I(\text{trust},q^*,3) = (1-x)(1-q^*)\beta h^I$, while $k_4^*\,(1-x+q^*x)\,\beta h^I = q^*\beta h^I$ (using (12)) experience an honest response, and thus a successfully completed transaction. In addition to the latter, also the trusting initiators in surviving matches experience a successfully completed transaction, resulting in $\pi^I(\text{trust},q^*,4) = 1-h^I+\beta q^* h^I$. Using (15), it then follows from (23) and (24) that

$$\widehat{\Pi}^I_{trusting}(m,q^*) = -\frac{x(1-q^*)\,(1-\alpha)\,\beta}{1-\alpha+\alpha\beta q^*}m - \frac{(1-x)(1-q^*)\,(1-\alpha)\,\beta}{1-\alpha+\alpha\beta q^*} + \frac{\beta q^*}{1-\alpha+\alpha\beta q^*}a, \tag{27}$$

and

$$\Pi^I_{trusting}(q^*) = -\frac{(1-x)(1-q^*)\,(1-\alpha)\,\beta}{1-\alpha+\alpha\beta q^*} + \frac{\beta q^*}{1-\alpha+\alpha\beta q^*}a.^{12} \tag{28}$$

All the mistrusting initiators are unmatched at the beginning of a typical period. The measure $1-\beta$ of them remain unmatched throughout the period, giving $\pi^I(\text{mistrust},q^*,1) = 1-\beta$. The remaining measure β of them become matched in a new match, but none of them initiate a transaction, even if they face a respondent with a clean record. Therefore $\pi^I(\text{mistrust},q^*,2) = \beta$. No other outcomes occur for the mistrusting initiators, giving $\pi^I(\text{mistrust},q^*,3) = \pi^I(\text{mistrust},q^*,4) = 0$. It then follows from (23) and (24) that

$$\widehat{\Pi}^I_{mistrusting}(m,q^*) = -\beta m, \tag{29}$$

[12]Here, and subsequently in all pecuniary payoff functions, we omit the redundant argument m (or later d).

and

$$\Pi^{I}_{mistrusting}(q^*) = 0. \tag{30}$$

Among the trustworthy respondents, the measure h^R are unmatched at the beginning of a typical period. Of these, βh^R become matched in a new match, while $(1-\beta)h^R$ remain unmatched throughout the period. Therefore π^R(trustworthy,$s_3^*, 1) = (1-\beta)h^R$. All of the βh^R trustworthy respondents matched in a new match have a clean record, putting their respective initiators into state I3. As a result, $(1-s_3^*)\,\beta h^R$ of these respondents do not experience an initiated transaction, giving π^R(trustworthy,$s_3^*, 2) = (1-s_3^*)\,\beta h^R$. The remaining $s_3^*\beta h^R$ of these respondents experience an initiated transaction, to which they respond honestly. In addition to the latter, also the trustworthy respondents in surviving matches experience a successfully completed transaction. Therefore π^R(trustworthy,$s_3^*, 4) = 1 - h^R + s_3^*\beta h^R$. No trustworthy respondents ever experience an outcome with a dishonest response, implying π^R(trustworthy,$s_3^*, 3) = 0$. Using (17), it then follows from (25) and (26) that

$$\widehat{\Pi}^{R}_{trustworthy}(d, s_3^*) = \Pi^{R}_{trustworthy}(s_3^*) = \frac{\beta s_3^*}{1-\alpha+\alpha\beta s_3^*}a. \tag{31}$$

Finally, all the untrustworthy respondents are unmatched at the beginning of a typical period. The measure $1-\beta$ of them remain unmatched throughout the period, implying π^R(untrustworthy,$s_3^*, 1) = 1-\beta$. The remaining measure β of them become matched in a new match. Of these respondents, $(1-x)\beta$ have a clean record, putting their respective initiators into state I3, while $x\beta$ have a spotty record, putting their initiators into state I2. As a result, $(1-s_3^*)\,(1-x)\beta$ of the respondents in the former group and all of the respondents in the latter group do not experience an initiated transaction, giving π^R(untrustworthy,$s_3^*, 2) = (1-s_3^*+s_3^*x)\,\beta$. The remaining $s_3^*(1-x)\beta$ respondents with a clean record experience an initiated transaction, to which they respond dishonestly, yielding π^R(untrustworthy,$s_3^*, 3) = s_3^*(1-x)\beta$. No mistrusting ini-

tiators ever experience the outcome with a successfully completed transaction, giving $\pi^R(\text{untrustworthy},s_3^*,4)=0$. It then follows from (25) and (26) that

$$\widehat{\Pi}^R_{untrustworthy}(d,s_3^*)=\beta s_3^*(1-x)(1-d), \tag{32}$$

and

$$\Pi^R_{untrustworthy}(s_3^*)=\beta s_3^*(1-x).^{13} \tag{33}$$

5.2 Can Anyone Gain by Committing to an Alternative Action?

Having calculated the expected total utility payoff of each individual given his or her actual behavior (trusting, mistrusting, being trustworthy, being untrustworthy), we can compare it to the expected total utility payoff in a counterfactual state of the world where he or she adopts the opposite action in state I3 or R4. This comparison will then allow us to evaluate whether any initiators or respondents could gain in equilibrium by committing *ex ante*, i.e., before knowing which state they are going to be in, to an alternative course of action in state I3 or state R4, respectively. The following lemma is a straightforward implication of (27), (29), (31), and (32).

Lemma 8.

$$\widehat{\Pi}^I_{trusting}(m,q^*)\gtreqless\widehat{\Pi}^I_{mistrusting}(m,q^*)\ \textit{as}\ m\gtreqless\widehat{m}(q^*,1)$$

[13] At this point the reader may wonder whether our results for the expected per period total utility payoffs are consistent with the expected discounted values, where the expectation is taken with respect to the equilibrium probability distribution of being in various states. Indeed, using the latter procedure yields the same expressions for the expected utility payoffs as presented in (27), (29), (31), and (32), except that they are multiplied by the factor $1/(1-\delta)$ that reflects discounting into the infinite future. We selected the per period expected utility approach for its simplicity and its ability to treat the expected pecuniary payoffs (for which no value functions are available from the previous analysis) analogously.

for any given m*, and*

$$\widehat{\Pi}^R_{trustworthy}(d, s_3^*) = \widehat{\Pi}^R_{untrustworthy}(d, s_3^*) \textit{ if } s_3^* = 0,$$
$$\widehat{\Pi}^R_{trustworthy}(d, s_3^*) > \widehat{\Pi}^R_{untrustworthy}(d, s_3^*) \textit{ if } x = 1,$$
$$\widehat{\Pi}^R_{trustworthy}(d, s_3^*) \gtreqless \widehat{\Pi}^R_{untrustworthy}(d, s_3^*) \textit{ as } d \gtreqless d(s_3^*, 1) \textit{ otherwise}$$

for any given d*.*[14]

This lemma implies that there is a threshold, given by $\widehat{m}(q^*, 1)$, for a behavioral predisposition towards trust above which initiators obtain a higher expected total utility payoff by trusting rather than not trusting, and vice versa below the threshold. Note, however, that since $\widehat{m}(q^*, \delta)$ is strictly decreasing in δ unless $q^* = 0$, when $\widehat{m}(q^*, \delta) = 1$ for all $\delta \in [0, 1]$ (and hence all the initiators are mistrusting), or $\beta = q^* = 1$, or $\beta = x = 1$, when $\widehat{m}(q^*, \delta) = -a$ for all $\delta \in [0, 1]$ (and hence all the initiators are trusting), this threshold is generically below the threshold $\widehat{m}(q^*, \delta)$ separating trusting from mistrusting initiators. As a result, there are generically mistrusting initiators who would be better off *ex ante* if they could commit to trusting. Note, however, that this result does *not* imply that these mistrusting initiators behave in a suboptimal way; on the contrary, they behave optimally given the particular state they are in. But because of their mistrusting behavior, compared to trusting initiators they find themselves more often in states that do not allow them to reap any gains from a successfully completed transaction. As a result, if they could choose between being trusting and mistrusting (and therefore between the probability distributions over the states of these two groups) before participating in the game, they would be better off forcing themselves to be trusting.

The discrepancy between the thresholds $\widehat{m}(q^*, 1)$ and $\widehat{m}(q^*, \delta)$ is due to the fact that

[14] Note that when $x = 1$,

$$\widehat{m}(q^*, \delta) = -\frac{a}{1 - \alpha\delta + \alpha\beta\delta} < -a < \underline{m},$$

and hence by Lemma 6 all the initiators trust in state I3, yielding $s_3^* = 1$. Therefore the cases $s_3^* = 0$ and $x = 1$ are mutually exclusive.

$\delta < 1$. As δ approaches unity, the gap between the two thresholds shrinks. In a sense, when δ is low, the initiators with m's between the two thresholds, when in state I3, are more preoccupied with the present danger of trusting rather than with the potential long-term benefit of starting a mutual cooperation, and hence they do not trust. However, as the discount factor δ is gradually increased towards unity, the future becomes more, and eventually overwhelmingly, important relative to the present. Since the future involves being in the four outcome situations with relative long-term frequencies equal to $\pi^I(m, i)$, $i \in \{1, 2, 3, 4\}$, respectively, it follows that as δ approaches unity, the comparison of discounted payoffs to trusting and not trusting in state I3 approaches the comparison of average per period total utility payoffs, and hence more and more initiators act in state I3 in a way that maximizes their *ex ante* expected total utility payoff.

Similarly, if $s_3^* > 0$ and $x < 1$, there is a threshold, given by $d(s_3^*, 1)$, for a behavioral predisposition towards trustworthiness above which respondents obtain a higher expected total utility payoff by being trustworthy rather than untrustworthy, and vice versa below the threshold. Again, as with the initiators, since $d(s_3^*, \delta)$ is strictly decreasing in δ, this threshold is below the threshold $d(s_3^*, \delta)$ separating trustworthy and untrustworthy respondents. As a result, when $s_3^* > 0$ and $x < 1$, there are untrustworthy respondents who would be better off *ex ante* if they could commit to being trustworthy. However, following the same intuition as with the initiators, as δ approaches unity, the gap between the two thresholds shrinks and hence more and more respondents act in state R4 in a way that maximizes their *ex ante* expected total utility payoff. When $s_3^* = 0$, there are no initiated transactions, implying that respondents' behavior in state R4 is irrelevant for their payoffs. As a result, the expected total utility payoffs of trustworthy and untrustworthy respondents coincide. When $x = 1$, untrustworthy respondents never get an opportunity to cheat, and hence always have a zero utility payoff. On the other hand, trustworthy respondents, if matched (which happens at least some of the time), always face an initiated transaction (see footnote 14), earning a positive utility payoff. As a result, trustworthy

respondents have a strictly higher expected total utility payoff.

These results are formally summarized in the following propositions:

Proposition 3. *When $q^* = 0$, all the initiators are mistusting and all of them are strictly better off* ex ante *by doing so. When $\beta = q^* = 1$ or $\beta = x = 1$, all the initiators are trusting and all of them are strictly better off* ex ante *by doing so. Otherwise every mistrusting initiator with $m \in (\widehat{m}(q^*, 1), \widehat{m}(q^*, \delta))$ would be strictly better off* ex ante *if she could commit to trusting, the mistrusting initiators with $m = \widehat{m}(q^*, 1)$ are equally well off* ex ante *by trusting or mistrusting, and all the other initiators are strictly better off* ex ante *by following their actual action in state I3.*

Proposition 4. *If $s_3^* = 0$, all the respondents are equally well off* ex ante *by being trustworthy or untrustworthy. Finally, if $x = 1$, all the untrustworthy respondents would be strictly better off* ex ante *if they could commit to being trustworthy. If $x < 1$ and $s_3^* > 0$, then every untrustworthy respondent with $d \in \left(\widehat{d}(s_3^*, 1), \widehat{d}(s_3^*, \delta)\right)$ would be strictly better off* ex ante *if he could commit to being trustworthy, the untrustworthy respondents with $d = \widehat{d}(s_3^*, 1)$ are equally well off* ex ante *by being trustworthy or untrustworthy, and all the other respondents are strictly better off following their actual action in state R4.*

5.3 Are Multiple Equilibria Pareto Ranked?

The first observation we make is that $\widehat{\Pi}^I_{trusting}(m, q^*)$ is strictly increasing in q^* (because $-a < \underline{m}$) and $\widehat{\Pi}^R_{trustworthy}(d, s_3^*)$ is strictly increasing in s_3^*. As a result, Example 2 of the Appendix illustrates that in general multiple equilibria are not Pareto ranked. In that example, respondents who respond honestly in all three equilibria are better off the higher the equilibrium level of trust s_3^*. On the contrary, initiators who trust in all three equilibria are better off the higher is the equilibrium level of trustworthiness q^*, and hence worse off the higher the equilibrium level of trust s_3^*.

In contrast, when the equilibrium levels of trust and trustworthiness are positively

related across various equilibria as in Example 1 in the Appendix, the equilibria are indeed Pareto ranked, with equilibria with a higher equilibrium level of trust and trustworthiness being Pareto superior. This result is formally established by the following proposition (see the Appendix for a proof):

Proposition 5. *Suppose that (s_3^A, q^A) and (s_3^B, q^B) are the values of (s_3, q) in two different equilibria, labelled A and B, respectively, such that $s_3^A < s_3^B$ and $q^A < q^B$. Then*

(a) all the initiators who trust in equilibrium A also trust in equilibrium B,

(b) all the respondents who are trustworthy in equilibrium A are also trustworthy in equilibrium B,

(c) the initiators who trust in equilibrium B are strictly better off in equilibrium B,

(d) the initiators who do not trust in equilibrium B (and hence do not trust in equilibrium A either) are equally well off in both equilibria, and

(e) all the respondents are strictly better off in equilibrium B.

As a result, equilibrium B Pareto dominates equilibrium A.

5.4 Do Trust and Trustworthiness Pay Off?

In this subsection, we are primarily interested in understanding how the comparative pecuniary payoff to trust, measured by $\Pi^I_{trusting}(q^*) - \Pi^I_{mistrusting}(q^*)$, relates to the trustworthiness of the respondents, and how the comparative pecuniary payoff to trustworthiness, measured by $\Pi^R_{trustworthy}(s_3^*) - \Pi^R_{untrustworthy}(s_3^*)$, relates to the trust of the initiators. Using (28) and (30), it follows that

$$\Pi^I_{trusting}(q^*) - \Pi^I_{mistrusting}(q^*) = -\frac{(1-x)(1-q^*)\,(1-\alpha)\,\beta}{1-\alpha+\alpha\beta q^*} + \frac{\beta q^*}{1-\alpha+\alpha\beta q^*}a,$$

which is strictly increasing in q^*. This invites the following intuitive interpretation: the comparative pecuniary payoff to trust increases with the trustworthiness of the respondents. However, a caveat applies here. Since q^* is an endogenous variable, an increase in q^* is usually associated with a change in some parameter, and that parameter may itself have a direct impact on the comparative pecuniary payoff to trust. Therefore the intuitive interpretation is valid only if one compares across multiple equilibria under one set of parameter values, or if a change in q^* is due to a change in δ, F, or G (the three parameters not directly affecting $\Pi^I_{trusting}(q^*) - \Pi^I_{mistrusting}(q^*)$). This result is formalized in the following proposition:

Proposition 6. *If comparing across multiple equilibria under one set of parameter values, or across equilibria that vary because of differences in δ, F, or G, the comparative pecuniary payoff to trust $\Pi^I_{trusting}(q^*) - \Pi^I_{mistrusting}(q^*)$ is strictly increasing in q^*.*

In a similar way, using (31) and (33), it follows that

$$\Pi^R_{trustworthy}(s_3^*) - \Pi^R_{untrustworthy}(s_3^*) = \frac{\beta s_3^*}{1-\alpha+\alpha\beta s_3^*}a - \beta s_3^*(1-x).$$

Whether this comparative pecuniary payoff to trustworthiness increases or decreases with s_3^* depends on the parameter values and the value of s_3^*. As above, a similar interpretation caveat applies. The analytical results for this case are summarized in the following proposition (the proof is omitted):

Proposition 7. *If comparing across multiple equilibria under one set of parameter values, or across equilibria that vary because of differences in δ, F, or G, the comparative pecuniary payoff to trustworthiness $\Pi^R_{trustworthy}(s_3^*) - \Pi^R_{untrustworthy}(s_3^*)$ is*

(a) strictly increasing in s_3^* over the entire domain $[0,1]$ if

$$x \geq 1 - \frac{1-\alpha}{(1-\alpha+\alpha\beta)^2}a,$$

(b) strictly decreasing in s_3^* over the entire domain $[0,1]$ if

$$x \leq 1 - \frac{a}{1-\alpha},$$

(c) strictly increasing in s_3^* over $[0,\overline{s}]$ and strictly decreasing in s_3^* over $(\overline{s},1]$ if

$$1 - \frac{a}{1-\alpha} < x < 1 - \frac{1-\alpha}{(1-\alpha+\alpha\beta)^2}a,$$

where

$$\overline{s} \equiv \frac{\left[\frac{(1-\alpha)a}{1-x}\right]^{\frac{1}{2}} - (1-\alpha)}{\alpha\beta}.$$

6 Conclusion

Empirical research has established that countries with a high proportion of trusting citizens tend to have a higher per capita income and to grow faster. What had not been established is the incentive people have to act in a trusting and trustworthy manner. In this paper we address this issue by developing an equilibrium matching model of trust and trustworthiness when individuals differ in their behavioral predispositions towards trusting and trustworthy behavior. We characterize how trust and trustworthiness impact each other, and how these interactions are affected by the observability of individuals' past behavior. We then combine these interactions in forming a general trust/trustworthiness equilibrium.

Our model unifies various partial approaches pursued in the previous literature. Its generality comes at the cost of ambiguous comparative statics results, mostly due to unrestricted forms of the distributions of the behavioral predispositions to trusting and trustworthy behavior. Nevertheless, it holds promise for providing a general conceptual framework for future empirical analyses of the complex relationship among trust, trust-

worthiness, and the prosperity of individuals and nations.

References

Alesina, Alberto and Eliana La Ferrara. Who Trusts Others? *Journal of Public Economics* 85(2): 207-234, 2002.

Arrow, Kenneth. Gifts and Exchanges. *Philosophy and Public Affairs* 1(4): 343-362, 1972.

Ashraf, Nava, Iris Bohnet, and Nikita Piankov. Decomposing Trust and Trustworthiness. *Experimental Economics* 9(3): 193-208, 2006.

Axelrod, Robert. An Evolutionary Approach to Norms. *American Political Science Review* 80(4): 1095-1111, 1986.

Berg, Joyce, John Dickhaut, and Kevin McCabe. Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10(2): 290-307, 1995.

Blonski, Matthias and Daniel Probst. The Emergence of Trust. Mannheim University, mimeo, 2001.

Chen, Yongmin. Promises, Trust, and Contracts. *Journal of Law, Economics, & Organization* 16(1): 209-232, 2000.

Coleman, James. *Foundations of Social Theory*. Cambridge, Massachusetts: Harvard University Press, 1990.

Dixit, Avinash. On Modes of Economic Governance. *Econometrica* 71(2): 449-481, 2003.

Fukuyama, Francis. *Trust.* New York: Basic Books, 1995.

Ghosh, Parikshit and Debraj Ray. Cooperation in Community Interaction Without Information Flows. *Review of Economic Studies* 63(3): 491-519, 1996.

Glaeser, Edward, David Laibson, Jose Scheinkman, and Christine Soutter. Measuring Trust. *Quarterly Journal of Economics* 115(3): 811-46, 2000.

Kandori, Michihiro. Social Norms and Community Enforcement. *Review of Economic Studies* 59(1): 63-80.

Knack, Stephen and Philip Keefer. Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *Quarterly Journal of Economics* 112(4): 1251-88, 1997.

Putnam, Robert (with R. Leonardi and R. Y. Nanetti). *Making Democracy Work*. Princeton: Princeton University, 1993. Press.

Rosenthal, R. Sequences of Games with Varying Opponents. *Econometrica* 47(6): 1353-1366, 1979.

Slemrod, Joel and Peter Katuscak. Do Trust and Trustworthiness Pay Off?, *Journal of Human Resources* 40(3): 621-646, 2004.

Sobel, Joel. A Theory of Credibility. *Review of Economic Studies* 52(4):557-573, 1985.

Tirole, Jean. A Theory of Collective Reputations (with applications to the persistence of corruption and to firm quality). *Review of Economic Studies* 63(1): 1-22, 1996.

Watson, Joel. Starting Small and Renegotiation. *Journal of Economic Theory* 85(1): 52-90.

Zak, Paul and Stephen Knack. Trust and Growth. *Economic Journal* 111(470): 295-321, 2001.

7 Appendix

Proof of Lemma 1. Consider a respondent in state R5. Given that this respondent has achieved this state, he must have responded honestly in the first period of the current match, i.e., in state R3 or R4. Equation (9), however, implies that any given respondent behaves identically in states R3 and R4. It must therefore be the case that the respondent's strategy for states R3 and R4 is to respond honestly, which, using (9), implies that

$$\begin{aligned} R_4(d) &= a + \alpha\delta\left[s_4 R_5(d) + (1-s_4)R_2(d)\right] + (1-\alpha)R_2(d) \\ &\geq 1 - d + R_1(d). \end{aligned} \tag{34}$$

In addition, (7) implies that

$$R_2(d) = \frac{\beta\delta s_3}{1-\delta+\beta\delta s_3} R_4(d). \tag{35}$$

Since it is always possible to avoid negative payoffs by behaving honestly and it is possible to earn a current positive payoff when in state R4, it must be the case that $R_4(d) > 0$. Then, because $\beta\delta s_3 / (1-\delta+\beta\delta s_3) \leq \delta$, it follows that

$$R_2(d) \leq \delta R_4(d). \tag{36}$$

Also (10) implies that

$$R_5(d) \geq \frac{a}{1-\alpha\delta} + \frac{1-\alpha}{1-\alpha\delta} R_2(d). \tag{37}$$

Next, we are going to show that $R_5(d) > R_2(d)$. Suppose, by contradiction, that $R_2(d) \geq R_5(d)$. Then (34) implies that

$$\begin{aligned} R_4(d) &\leq a + \alpha\delta R_2(d) + (1-\alpha)R_2(d) \\ &= a + (1 - \alpha + \alpha\delta)\, R_2(d), \end{aligned}$$

and hence

$$\delta R_4(d) \leq \delta a + \delta\left(1 - \alpha + \alpha\delta\right) R_2(d).$$

Combining this result with (36) then gives

$$R_2(d) \leq \delta a + \delta\left(1 - \alpha + \alpha\delta\right) R_2(d),$$

which is equivalent to

$$R_2(d) \leq \frac{\delta a}{1 - \delta\left(1 - \alpha + \alpha\delta\right)}. \tag{38}$$

In addition, if $R_2(d) \geq R_5(d)$, then (37) implies that

$$R_5(d) \geq \frac{a}{1 - \alpha\delta} + \frac{1 - \alpha}{1 - \alpha\delta} R_5(d). \tag{39}$$

If $\alpha = 0$, it follows from this result that $0 \geq a$, which contradicts the assumption $a > 0$. Therefore if (39) holds, it must be the case that $\alpha > 0$. Then (39) is equivalent to

$$R_5(d) \geq \frac{a}{\alpha(1 - \delta)}. \tag{40}$$

However, because $a > 0$ and

$$\frac{\delta}{1 - \delta\left(1 - \alpha + \alpha\delta\right)} < \frac{1}{\alpha(1 - \delta)},$$

(38) and (40) imply that $R_5(d) > R_2(d)$, which is in contradiction to the starting assumption that $R_2(d) \geq R_5(d)$. Therefore it must be the case that $R_5(d) > R_2(d)$.

This result, combined with (34), gives

$$\begin{aligned} a + \alpha\delta R_5(d) + (1 - \alpha) R_2(d) &\geq a + \alpha\delta\left[s_4 R_5(d) + (1 - s_4) R_2(d)\right] + (1 - \alpha) R_2(d) \\ &\geq 1 - d + R_1(d). \end{aligned}$$

Using (10) and Assumption 2, this implies that the respondent prefers to respond honestly in state R5. □

Proof of Lemma 5. Consider an initiator in state I4. Given that this initiator has achieved this state, she must have initiated in the first period of the current match. By Lemma 4, no initiator initiates in state I2. Therefore in the first period of the current match the initiator must have been in state I3. This implies that her strategy for state I3 is to initiate, which, using (3), implies that

$$\begin{aligned} I_3(m) &= k_4 \left[a + \alpha\delta I_4(m) + (1-\alpha) I_1(m)\right] + (1-k_4)\left[-1 + I_1(m)\right] \\ &\geq -m + I_1(m). \end{aligned} \tag{41}$$

Recall that nobody is a pathological truster ($\overline{m} < 1$). This has two implications. First, it must be the case that $k_4 > 0$, since otherwise (41) would imply that $1 \leq m$. Intuitively, because by Lemma 3 $k_3 = 0$, no initiator would ever be in state I4 if $k_4 = 0$. Second, $-1 + I_1(m) < -m + I_1(m)$. Using these two implications, (41) implies that

$$a + \alpha\delta I_4(m) + (1-\alpha) I_1(m) \geq -m + I_1(m). \tag{42}$$

Next, we are going to show that $I_5(m) \geq I_4(m)$. Suppose, by contradiction, that $I_4(m) > I_5(m)$. Because, using Lemma 1, (4) and (5) yield

$$I_4(m) = \max\left\{-m + I_1(m); I_5(m)\right\},$$

the inequality $I_4(m) > I_5(m)$ implies that $I_4(m) = -m + I_1(m)$. Using this result to substitute for the right-hand side in (42) and rearranging gives

$$I_4(m) \leq \frac{a}{1-\alpha\delta} + \frac{1-\alpha}{1-\alpha\delta} I_1(m).$$

However, (5) yields

$$I_5(m) = \frac{a}{1-\alpha\delta} + \frac{1-\alpha}{1-\alpha\delta} I_1(m),$$

which implies that $I_5(m) \geq I_4(m)$, which is a contradiction. Therefore it must be the case that $I_5(m) \geq I_4(m)$.

This result, combined with (42), then gives

$$a + \alpha\delta I_5(m) + (1-\alpha) I_1(m) \geq -m + I_1(m),$$

and hence, using (4) and Lemma 1, the initiator prefers to initiate in state I4. □

Proof of Lemma 6. Combining Lemmata 5 and 1 with (4) and (5) implies that

$$I_4(m) = I_5(m) = \frac{a}{1-\alpha\delta} + \frac{1-\alpha}{1-\alpha\delta} I_1(m). \tag{43}$$

Using this result for substitution into (3) gives

$$I_3(m) = \max\left\{-m + I_1(m); \frac{k_4 a}{1-\alpha\delta} + k_4 - 1 + \left[1 - \frac{k_4\alpha(1-\delta)}{1-\alpha\delta}\right] I_1(m)\right\}. \tag{44}$$

In addition, substituting from (11) to (1) and rearranging gives

$$I_1(m) = -\frac{\beta\delta x(1-q)}{1-\delta+\beta\delta(1-x+qx)} m + \frac{\beta\delta(1-x+qx)}{1-\delta+\beta\delta(1-x+qx)} I_3(m). \tag{45}$$

Finally, substituting (45) into (44) gives

$$\begin{aligned} I_3(m) = \max\Bigg\{ & -\frac{1-\delta+\beta\delta}{1-\delta} m; \frac{1-\delta+\beta\delta(1-x+qx)}{1-\delta}\left(\frac{k_4 a}{1-\alpha\delta} + k_4 - 1\right) \\ & + \frac{\beta\delta x(1-q)\left[k_4\alpha(1-\delta) - (1-\alpha\delta)\right]}{(1-\delta)(1-\alpha\delta)} m - \frac{k_4\alpha\beta\delta(1-x+qx)}{1-\alpha\delta} I_3(m)\Bigg\}. \end{aligned} \tag{46}$$

If not trusting is the preferred choice in state I3, then by (46) it must be the case that

$$\begin{aligned} I_3(m) &= -\frac{1-\delta+\beta\delta}{1-\delta}m \\ &> \frac{1-\delta+\beta\delta(1-x+qx)}{1-\delta}\left(\frac{k_4 a}{1-\alpha\delta}+k_4-1\right) \\ &+ \frac{\beta\delta x(1-q)\left[k_4\alpha(1-\delta)-(1-\alpha\delta)\right]}{(1-\delta)(1-\alpha\delta)}m - \frac{k_4\alpha\beta\delta(1-x+qx)}{1-\alpha\delta}I_3(m), \end{aligned} \tag{47}$$

which implies that

$$m < \frac{(1-k_4)(1-\alpha\delta)-ak_4}{1-\alpha\delta+\alpha\beta\delta k_4}. \tag{48}$$

In summary, if not trusting is the preferred choice in state I3, then (48) holds. Then, by contrapositive, if

$$m \geq \frac{(1-k_4)(1-\alpha\delta)-ak_4}{1-\alpha\delta+\alpha\beta\delta k_4},$$

trusting is the preferred choice in state I3.

If trusting is the preferred choice in state I3, then by (46) it must be the case that

$$\begin{aligned} I_3(m) &= \frac{1-\delta+\beta\delta(1-x+qx)}{1-\delta}\left(\frac{k_4 a}{1-\alpha\delta}+k_4-1\right) \\ &+ \frac{\beta\delta x(1-q)\left[k_4\alpha(1-\delta)-(1-\alpha\delta)\right]}{(1-\delta)(1-\alpha\delta)}m - \frac{k_4\alpha\beta\delta(1-x+qx)}{1-\alpha\delta}I_3(m) \\ &\geq -\frac{1-\delta+\beta\delta}{1-\delta}m, \end{aligned} \tag{49}$$

which implies that

$$\begin{aligned} I_3(m) &= \frac{\left[1-\delta+\beta\delta\left(1-x+qx\right)\right]\left[k_4 a-(1-k_4)(1-\alpha\delta)\right]}{(1-\delta)\left[1-\alpha\delta+\alpha\beta\delta k_4\left(1-x+qx\right)\right]} \\ &+ \frac{\beta\delta x(1-q)\left[k_4\alpha(1-\delta)-(1-\alpha\delta)\right]}{(1-\delta)\left[1-\alpha\delta+\alpha\beta\delta k_4\left(1-x+qx\right)\right]}m, \end{aligned} \tag{50}$$

and

$$m \geq \frac{(1-k_4)(1-\alpha\delta)-ak_4}{1-\alpha\delta+\alpha\beta\delta k_4}. \tag{51}$$

In summary, if trusting is the preferred choice in state I3, then (51) holds. Then, by contrapositive, if

$$m < \frac{(1-k_4)(1-\alpha\delta) - ak_4}{1-\alpha\delta+\alpha\beta\delta k_4},$$

not trusting is the preferred choice in state I3.

The final result then follows by noting that, using (12),

$$\frac{(1-k_4)(1-\alpha\delta) - ak_4}{1-\alpha\delta+\alpha\beta\delta k_4} = \frac{(1-\alpha\delta)(1-x)(1-q) - aq}{(1-\alpha\delta)(1-x+qx)+\alpha\beta\delta q}.$$

□

Proof of Lemma 7. Using Lemma 4, (6) implies that

$$R_1(d) = \frac{\beta\delta s_3(1-x)}{1-\delta+\beta\delta s_3(1-x)} R_4(d).$$

Similarly, using Lemma 1, (10) implies that

$$R_5(d) = \frac{a}{1-\alpha\delta} + \frac{1-\alpha}{1-\alpha\delta} R_2(d).$$

Using these two results, (35), and Lemma 5 for substitution into (9), we obtain

$$\begin{aligned} R_4(d) = \max\Bigg\{ & 1-d+\frac{\beta\delta s_3(1-x)}{1-\delta+\beta\delta s_3(1-x)} R_4(d); \frac{a}{1-\alpha\delta} \\ & + \frac{(1-\alpha)\,\beta\delta s_3}{(1-\alpha\delta)\,(1-\delta+\beta\delta s_3)} R_4(d) \Bigg\}. \end{aligned} \tag{52}$$

If being untrustworthy is the preferred choice in state R4, then by (52) it must be the case that

$$\begin{aligned} R_4(d) &= 1-d+\frac{\beta\delta s_3(1-x)}{1-\delta+\beta\delta s_3(1-x)} R_4(d) \\ &> \frac{a}{1-\alpha\delta} + \frac{(1-\alpha)\,\beta\delta s_3}{(1-\alpha\delta)\,(1-\delta+\beta\delta s_3)} R_4(d), \end{aligned} \tag{53}$$

which implies

$$R_4(d) = \frac{1-\delta+\beta\delta s_3(1-x)}{1-\delta}\,(1-d)\,, \tag{54}$$

and

$$d < 1 - \frac{1-\delta+\beta\delta s_3}{(1-\alpha\delta+\alpha\beta\delta s_3)\,[1-\delta+\beta\delta s_3(1-x)]}a. \tag{55}$$

In summary, if being untrustworthy is the preferred choice in state R4, then (55) holds. Then, by contrapositive, if

$$d \geq 1 - \frac{1-\delta+\beta\delta s_3}{(1-\alpha\delta+\alpha\beta\delta s_3)\,[1-\delta+\beta\delta s_3(1-x)]}a,$$

being trustworthy is the preferred choice in state R4.

If being trustworthy is the preferred choice in state R4, then by (52) it must be the case that

$$\begin{aligned} R_4(d) &= \frac{a}{1-\alpha\delta} + \frac{(1-\alpha)\,\beta\delta s_3}{(1-\alpha\delta)\,(1-\delta+\beta\delta s_3)}R_4(d) \\ &\geq 1 - d + \frac{\beta\delta s_3(1-x)}{1-\delta+\beta\delta s_3(1-x)}R_4(d), \end{aligned} \tag{56}$$

which implies

$$R_4(d) = \frac{1-\delta+\beta\delta s_3}{(1-\delta)\,(1-\alpha\delta+\alpha\beta\delta s_3)}a, \tag{57}$$

and

$$d \geq 1 - \frac{1-\delta+\beta\delta s_3}{(1-\alpha\delta+\alpha\beta\delta s_3)\,[1-\delta+\beta\delta s_3(1-x)]}a. \tag{58}$$

In summary, if being trustworthy is the preferred choice in state R4, then (58) holds. Then, by contrapositive, if

$$d < 1 - \frac{1-\delta+\beta\delta s_3}{(1-\alpha\delta+\alpha\beta\delta s_3)\,[1-\delta+\beta\delta s_3(1-x)]}a,$$

being untrustworthy is the preferred choice in state R4. □

Proof of Proposition 2. By Proposition 1, the question of the existence of a general trust/trustworthiness equilibrium is equivalent to the question of whether the map $h : [0,1]^2 \to [0,1]^2$ defined by $h^1(z) \equiv s(z_2)$ and $h^2(z) \equiv q(z_1)$ has a fixed point. Because $F(\cdot)$, $G(\cdot)$, $m(\cdot)$, and $d(\cdot)$ are all continuous, $h(\cdot)$ is a continuous function from a closed, bounded and convex subset of R^2 into itself, and so it follows by the Brouwer Fixed Point Theorem that it has a fixed point. Therefore a general trust/trustworthiness equilibrium exists. □

Example 1. *Let $a = 0.3$, $\alpha = 0$, $\beta = 0.5$, $\delta = 0.8$, $x = 0.5$, F be the Beta distribution with parameters $(6,6)$, and G be the Beta distribution with parameters $(5,5)$ scaled on the support $[0, 0.99]$. There are three equilibria. In the low trust/low trustworthiness equilibrium, $(s_3, q, k_4) = (0.017, 0.084, 0.1551)$, in the medium trust/ medium trustworthiness equilibrium $(s_3, q, k_4) = (0.4192, 0.222, 0.3633)$, and in the high trust/high trustworthiness equilibrium $(s_3, q, k_4) = (0.7991, 0.3246, 0.4901)$.*

Example 2. *Let $a = 0.15$, $\alpha = 0.99$, $\beta = 0.5$, $\delta = 0.9$, $x = 0$, F be the Beta distribution with parameters $(20, 50)$, and*

$$G(u) = \frac{2}{3}u\chi_{\{0\leq u<0.9\}} + [0.6 + 5(u - 0.9)]\chi_{\{0.9\leq u<0.98\}} + \chi_{\{0.98\leq u\}},$$

where $\chi_{(\cdot)}$ is the indicator function. There are three equilibria. In the low trust/high trustworthiness equilibrium, $(s_3, q, k_4) = (0.2275, 0.0692, 0.0692)$, in the medium trust/medium trustworthiness equilibrium $(s_3, q, k_4) = (0.2458, 0.0307, 0.0307)$, and in the high trust/low trustworthiness equilibrium $(s_3, q, k_4) = (0.2606, 0.0161, 0.0161)$.

Proof of Proposition 5. Since $m(q, \delta)$ is strictly decreasing in q, it follows that $m(q^A, \delta) > m(q^B, \delta)$, and hence by Lemma 6 all the initiators who trust in equilibrium A also trust in equilibrium B. Similarly, because $q^A < q^B$ and $s_3^A < s_3^B$, (18) implies that $F\left[d(s_3^A, \delta)\right] > F\left[d(s_3^B, \delta)\right]$, and hence $d(s_3^A, \delta) > d(s_3^B, \delta)$. Therefore by Lemma 7 all the respondents who are trustworthy in equilibrium A are also trustworthy in equilibrium B.

Also note that when $x = 1$, then there is a unique equilibrium with

$$(s_3, q, k_4) = \left(1, \frac{(1-\alpha)F(d^*)}{1-\alpha+\alpha\beta F(d^*)}, 1\right),$$

where

$$d^* = 1 - \frac{1-\delta+\beta\delta}{(1-\alpha\delta+\alpha\beta\delta)(1-\delta)}a.$$

Therefore the assumed multiplicity of equilibria implies that $x < 1$.

Having established these basics, first consider the initiators who prefer to trust in either equilibrium. Since $\widehat{\Pi}^I_{trusting}(m, q^*)$ is strictly increasing in q^*, these initiators are strictly better off in equilibrium B. Second, consider the initiators who prefer not to trust in equilibrium A, but prefer to trust in equilibrium B. Since $\widehat{\Pi}^I_{mistrusting}(m, q^*)$ is independent of q^*, not trusting in equilibrium A has the same payoff as not trusting in equilibrium B. In addition, since $q^B > q^A \geq 0$ and $x < 1$, Proposition 3 implies that since these initiators trust in equilibrium B, they have strictly higher payoffs than what they would get by not trusting in this equilibrium. As a result, these initiators are strictly better off in equilibrium B. Third, consider the initiators who prefer not to trust in either equilibrium. Since $\widehat{\Pi}^I_{mistrusting}(m, q^*)$ is independent of q^*, these initiators are equally well off in either equilibrium.

Let us now turn to the respondents. First, consider the respondents who prefer to be trustworthy in either equilibrium. Since $\widehat{\Pi}^R_{trustworthy}(d, s_3^*)$ is strictly increasing in s_3^*, these respondents are strictly better off in equilibrium B. Second, consider the respondents who prefer to be untrustworthy in equilibrium A, but prefer to be trustworthy in equilibrium B. Because these respondents are untrustworthy in equilibrium A, Lemma 7 implies that $d < 1$ for any of these respondents. Since $x < 1$, it then follows that for these respondents $\widehat{\Pi}^R_{untrustworthy}(d, s_3^*)$ is strictly increasing in s_3^*, and hence not trusting in equilibrium B has a strictly higher payoff than not trusting in equilibrium A. In addition, since $s_3^B > s_3^A \geq 0$, Proposition 4 implies that whoever behaves in a trustworthy

way in equilibrium B is strictly better off by doing so than by being untrustworthy. As a result, these respondents are strictly better off in equilibrium B. Third, consider the respondents who prefer to be untrustworthy in either equilibrium. By Lemma 7, for any of these respondents it must be the case that $d < 1$. Since $x < 1$, it then follows that for these respondents $\widehat{\Pi}^R_{untrustworthy}(d, s_3^*)$ is strictly increasing in s_3^*, and hence these respondents are strictly better off in equilibrium B. □

Individual researchers, as well as the on-line and printed versions of the CERGE-EI Working Papers (including their dissemination) were supported from the following institutional grants:

- Economic Aspects of EU and EMU Entry [Ekonomické aspekty vstupu do Evropské unie a Evropské měnové unie], No. AVOZ70850503, (2005-2010);
- Economic Impact of European Integration on the Czech Republic [Ekonomické dopady evropské integrace na ČR], No. MSM0021620846, (2005-2011);

Specific research support and/or other grants the researchers/publications benefited from are acknowledged at the beginning of the Paper.

Published by
Charles University in Prague, Center for Economic Research and Graduate Education (CERGE)
and
Economics Institute (EI), Academy of Sciences of the Czech Republic
CERGE-EI, Politických vězňů 7, 111 21 Prague 1, tel.: +420 224 005 153, Czech Republic.
Printed by CERGE-EI, Prague
Subscription: CERGE-EI homepage: http://www.cerge-ei.cz

Editors: Directors of CERGE and EI
Managing editors: Deputy Directors for Research of CERGE and EI

ISSN 1211-3298
ISBN 80-7343-101-7 (Univerzita Karlova. Centrum pro ekonomický výzkum a doktorské studium)
ISBN 80-7344-090-3 (Akademie věd České republiky. Národohospodářský ústav)

CERGE-EI
P.O.BOX 882
Politických vězňů 7
111 21 Praha 1
Czech Republic
http://www.cerge-ei.cz