A Game-Theoretic Framework for Analyzing Trust-Inference Protocols[*]

Ruggero Morselli[†] Jonathan Katz[†] Bobby Bhattacharjee[†]

Abstract

We propose a novel game-theoretic framework for analyzing the robustness of trust-inference protocols in the presence of adversarial (but rational) users. To the best of our knowledge, this is the first such framework which simultaneously (1) admits a rigorous and precise definition, thereby enabling formal proofs of security (in various adversarial settings) for specific trust-inference protocols; (2) is flexible enough to accommodate a full range of (realistic) adversarial behavior and network models; and (3) is appropriate for *decentralized* networks, and in particular does not posit a trusted, centralized party with complete knowledge of the system history. We also show some preliminary results regarding the design of trust-inference protocols which can be rigorously *proven secure* within our model.

In addition to establishing a solid foundation for future work, our framework also enables a rigorous and objective comparison among existing trust inference protocols.

1 Introduction

Peer-to-peer networks require a significant amount of cooperation among their members in order to fully realize their potential. For example, in a resource-sharing system where users trade, say, spare computer cycles, such cooperation is crucial to the functioning of the system. Indeed, if too many "free riders" (i.e., those who use system resources without donating any of their own) are present, the utility of the system as a whole — especially for "good" users who *do* donate their resources — will markedly decrease.

Enforcing such cooperation in a peer-to-peer network is, unfortunately, rather difficult, especially because there is no central authority with the ability to "punish" non-cooperating users. To address this issue, a large body of work has recently focused on providing *incentives* for players in the system to behave in a cooperative manner.[1] One straightforward way to attempt to enforce such cooperation is via a "payback" mechanism in which user A donates (some amount of) resources to user B only as long as B continues to donate (a similar amount of) resources to A. (I.e., A is willing to extend B only a certain amount of "credit" at which point B must begin paying back.) Such simple mechanisms are, however, rather limited. First, *direct* interaction between A and B may occur infrequently (or even only once!), giving A little or no chance to "redeem" resources donated to B. Second, it is unclear what happens if, say, B leaves the system before having had the chance to pay back A. A solution of this type also strongly biases users toward interacting *only* with parties with whom they have had direct previous (positive) experience. Although intuitively appealing, this potentially limits the overall utility of the system since users will tend to trade repeatedly with the same partners rather than explore new partners. We note that it also makes it more difficult for new users to be fully integrated into the system. Finally, it is not clear what happens when participants interact for the first time (i.e., when the system is first initialized): when A and B first meet, who first extends credit to whom?

The above drawbacks motivate the idea of having users base their future actions on *more* than just their own personal history of prior interactions; see, e.g., [10] for simulations and discussion further illustrating this point. In particular, one might hope to de-

[*] Full version — Draft. An abstract of this work appeared in Second Workshop on the Economics of Peer-to-Peer Systems, June 4–5 2004, Boston, MA, USA.

[†] {ruggero,jkatz,bobby}@cs.umd.edu. Department of Computer Science, University of Maryland. Work supported by NSF Trusted Computing Grant #0310499.

[1] We stress that the intention of such incentives is not to prevent attacks on the system by *malicious* users, but rather to enforce cooperative behavior in the network on the part of *rational* (self-interested) users. Purely malicious behavior must be handled using more traditional security measures.

sign *trust-inference*[2] mechanisms by which information about user behavior can be propagated throughout the system; in this way (at least in theory), a user A who freely donates resources will be rewarded as others will be more likely to share resources with A, while a user B who takes resources without giving any of his own will be punished to the extent that others refuse to share resources with B. A growing recognition of the importance of trust inference has led to an extensive amount of research focused on designing "good" trust-inference protocols, both in the specific context of peer-to-peer networks as well as in more general settings. We cannot survey all prior work here, but refer the reader to [6, 13, 17, 19, 20, 12, 1, 3, 2, 10, 14, 5, 18, 9] as representative examples.

Unfortunately, we are aware of very few works which rigorously define what a "good" trust-inference protocol should achieve! Instead, a lot of work in this area has been rather heuristic and ad-hoc, proposing solutions satisfying some list of properties but with no indication that these are the "right" properties one should aim for. In other cases, simulations or informal arguments indicate that a proposed trust-inference protocol is resilient to a *particular* adversarial strategy (or strategies), but no proof is offered to show that the protocol is resilient to *all* (rational) adversarial strategies. Finally, many works make unjustified assumptions; for example, some works assume that although users may cheat by refusing to share resources, all users honestly report the behavior of other nodes (or even their own behavior!).

Some notable exceptions (see [6, 5, 18]) provide a formal adversarial model, a definition of robustness[3], and proofs that a proposed protocol is robust under the given definition. However, all work of this type of which we are aware assumes some form of *global knowledge* which would be implemented in practice using some centralized mechanism. For example, Friedman and Resnick [6] posit that all parties have complete and accurate knowledge of the previous behavior of all other users in the system; other work focusing on the "E-bay model" (see [5]) assumes a public, incorruptible bulletin board on which users post feedback about each other.

1.1 Our Contribution

In the course of our ongoing work developing and analyzing protocols for trust inference in completely decentralized systems [11, 16], we have been frustrated by the lack a formal model in which to evaluate our proposed mechanisms, as well as the lack of any objective way to compare the robustness our protocols with previously-proposed ones. The framework we propose here was developed in response to this need, and we hope it will prove useful to other researchers in this area. We stress that the model presented here is very preliminary, but will hopefully serve as a basis and as an impetus for much-needed future work in this domain. (For those who do not agree with the particulars of our model, we hope they agree that *some* formal model is sorely needed!) As we see it, the advantages of our framework include:

- It admits a concrete and precise definition, thereby enabling rigorous proofs of security (in a chosen adversarial model) for specific protocols.
- Similarly, the definition enables an objective way to compare existing trust-inference protocols and to determine their suitability for various systems under a given adversarial model.
- Our definition assumes no "global knowledge", centralized infrastructure, or pre-provisioned trusted parties, and is therefore appropriate for modeling completely decentralized systems with no central authority. However, we note that our model may be easily augmented to include a trusted authority should one choose to do so.
- Our definition is *flexible* enough to allow consideration of a wide range of adversarial behavior and system models (such as adversarial coalitions, Sybil attacks [4], and asynchronous trading) not typically handled (in a formal way) by previous work.

In Section 2 we discuss our framework and define our notion of *robust* trust inference. We also discuss additional desiderata which provide ways of discriminating among robust trust-inference protocols. In Section 3 we give preliminary results indicating that robust protocols are achievable even in very strong adversarial environments (i.e., allowing for arbitrary-size coalitions, Sybil attacks, asynchronous trading, and easy-to-change pseudonyms) *without any centralized infrastructure*. We warn the reader that these results are only meant to illustrate the *feasibility* of realizing our definition of robustness; developing more

[2] These are also known in the literature as *trust-propagation protocols* or *reputation/recommender systems*. All of these seek to accomplish essentially the same task.

[3] Fixing the adversarial model is usually the difficult part, since a robust protocol is almost always defined as one whose actions form a game-theoretic equilibrium (i.e., an adversary has no reason to deviate from the prescribed actions).

efficient protocols (which remain provably robust) is the subject of ongoing research.

2 Adversarial Framework

We define our adversarial framework in two stages. First, we describe our basic framework which can be used to model essentially any sort of adversarial behavior and/or network in a very simple way. Jumping ahead a bit, we then define our notion of robustness which will remain unchanged even as the adversarial model is adjusted. (Basically, a trust-inference protocol is *robust* if the actions prescribed by the protocol form a game-theoretic equilibrium. We stress, however, that single-player deviations in our model actually correspond to adversarial *coalitions* in the real network.) Our basic framework gives the adversary a considerable amount of power, and is probably too pessimistic for modeling realistic threats in real-world systems.[4] Thus, we discuss a number of ways of extending our model (which have the effect of restricting the power of the adversary). Our goal here is to highlight the flexibility and generality of our approach, rather than to suggest any particular choice of adversarial model. Indeed, different adversarial models are better suited for different environments and so there is no "best" model to consider.

2.1 Basic Framework

A key component of our framework is the notion of a *pseudonym* by which a user is known to others in the network. We assume pseudonyms with the following properties: they are distinct, they are easy to generate by users *themselves* (and do not require the services of a trusted party), and it is impossible to impersonate another party by using their pseudonym. All these properties are (essentially) satisfied by identifying pseudonyms with public keys for a secure digital signature scheme [8].[5] We stress that these public keys are not assumed to be registered in any central location, and need not be certified in any way. In particular, although we assume that honest participants use the same pseudonym throughout their entire lifetime, an adversary can easily generate new pseudonyms as often as it likes.

Our model gives the adversary almost complete control of the system. For convenience, we use the standard conventions of the cryptographic community and model adversarial actions using various *oracles*. Some of these oracles correspond to actions of a real-world adversary, while others merely offer a convenient way of considering the worst-case scenario of events which (in the real world) are outside the adversary's control.[6] Given a trust-inference protocol Π (whose details are entirely known to the adversary), we provide adversary $\mathcal{A}$ the following oracles:

- $\mathsf{NewUser}$ creates a new honest user in the system, and $\mathcal{A}$ learns this user's pseudonym. A party using pseudonym i is simply called "user i".
- $\mathsf{HonestPlay}(i, j)$ causes honest users i and j to play an instance of some 2-player game (e.g., prisoners' dilemma). In playing this game, the users will behave exactly in accordance with protocol Π. Note that Π prescribes both how trust should be inferred as well as how a user's actions should depend upon the inferred value.
- $\mathsf{Play}(i, \mathit{id}, \mathit{action})$ plays a 2-player game between $\mathcal{A}$ (using pseudonym id) and honest player i. The adversary plays *action* while i behaves in accordance with Π. The adversary may not use an *id* held by an honest party (this would amount to impersonation, which is assumed impossible).
- $\mathsf{Send}(i, \mathit{id}, \mathit{msg})$ sends *msg* to honest player i "from" player *id*, where again we require that *id* not be held by an honest party. This models messages $\mathcal{A}$ sends as part of Π (of course, $\mathcal{A}$ need not behave according to Π).

We do not provide an oracle enabling honest players to send messages to each other; this is the one part of our model *not* under adversarial control. Instead, we assume that Π is executed faithfully among the honest users "in the background" and without any interference from $\mathcal{A}$ (except for messages $\mathcal{A}$ can send on behalf of pseudonyms it controls). We assume that $\mathcal{A}$ can see any messages sent between honest users as part of Π. (Note that if Π is deterministic, then $\mathcal{A}$ automatically knows these messages anyway.)

[4] Yet, it is interesting that robust trust-inference protocols exist even for our strongest adversarial model; see Section 3.

[5] We stress two caveats here: first, equating digital signatures with pseudonyms is only sound when considering *computationally-limited* (e.g., poly-time) adversaries, as is typically the case of interest. Second, the implicit assumption is that users will be careful not to "leak" the associated secret key. Maintaining secrecy of secret keys is a security concern that lies outside the game-theoretic framework considered here.

[6] If a protocol is robust even against an ideal adversary having this level of control over the network, then clearly it will also be robust against a real-world adversary.

For simplicity and concreteness, we assume (following [6]) that all 2-party games are a "prisoners' dilemma" with the payouts indicated below (where C represents "cooperate" and D represents "defect"):

	C	D
C	(1, 1)	(-1,2)
D	(2, -1)	(0, 0)

In particular, our results in Section 3 assume the payoff matrix above. Note, however, that our framework easily accommodates different games, payouts that change with time (or according to Π), or adversarial selection of the game to play.

Robustness. In order to model robustness in game-theoretic terms, we need to add a notion of time (as well as a *discount factor*) to our model. Here, we do so in a very general fashion (essentially giving the adversary the most power); we discuss some more restrictive ways of dealing with time below.

We assume that each time the adversary $\mathcal{A}$ makes an oracle call, it associates with the call a particular time t (where $t \geq 0$ is an integer). Other than the fact that the time t can never decrease, our only restriction is that $\mathcal{A}$ "can't do too much in too short a time"; thus, $\mathcal{A}$ can make at most N NewUser calls with the same value of t (i.e., only some bounded number of users join at any particular time) and at most N' Play calls with the same value of t (i.e., $\mathcal{A}$ cannot trade with too many people at the same time). Finally, Π is assumed to be run whenever the adversary "moves the clock forward". I.e., when $\mathcal{A}$ makes its first oracle call at some time t, we assume that the honest players run Π immediately beforehand, based on the events that have occurred up to time t. We stress that $\mathcal{A}$ may interact with multiple parties at some instant t without giving these parties any chance to run Π in the interim.

Note that one may always set N, N' as large as one likes, and thus the above does not fundamentally restrict the adversary's power. However, a given protocol may only be provably secure when N, N' are lower than a certain bound. (The implication is that the protocol is secure against one class of adversaries, but not necessarily secure when the adversary has more power: e.g., in case the adversary releases a virus giving it control over a huge number of hosts.)

We measure the *utility* of the adversary as follows. Each time the adversary makes a Play oracle call at some time t, the adversary's utility increases by $\delta^t \mu$, where μ is the payoff given by the matrix above (i.e., if $\mathcal{A}$ plays D and the honest user player C, then $\mu = 2$) and $\delta < 1$ is a *discount factor* [7]. We assume a rational adversary who wishes to maximize its total utility as time tends to infinity.

2.2 The Network Model

In this section we discuss the model of the network as seen by the adversary and by the honest users.

Our framework can be applied to different network models. Here we just describe two simple examples.

Broadcast Network. There exists a reliable broadcast channel.

Complete Point-to-Point Network. The network has a trusted infrastructured that offers the following services:

1. Every time a new (honest) user i joins the system, the infrastructure notifies every other user j of this fact; user j learns the identity i of the new user.
2. Any user j can use the infrastructure to send a message to user i.
3. The adversary is given an additional oracle NotifyJoin(i, j) that makes the infrastructure notify honest user j that a user with identity i (not currently held by any honest user) has joined the system.

2.3 The Timing Model

In this section we give a formal description of what we mean by saying that Π is assumed to be run whenever the adversary "moves the clock forward".

The assumptions are the following:

- Each time period t is divided into two disjoint phases: the first is called the *play phase* and the second the *protocol phase*.
- During the play phase the adversary can issue NewUser, Play and HonestPlay oracle calls.
- The play phase ends and the protocol phase begins when the adversary makes a Send or Activate (see below) oracle call with the same time stamp t as the current period.
- During the protocol phase the adversary can issue the Send oracle call. In addition, the adversary can issue an Activate(i) oracle call, where i is the identity of a honest user. Such oracle

call causes user i to receive all the messages that have been sent to it by honest users, perform a step of computation of the protocol and causes i either to perform an operation on the network infrastructure (for example: send a message m to some user j or send a message m to the broadcast channel) or to determine that no further action is needed. The call returns the action performed by the player in the former case or the special Done symbol in the latter case.

- The protocol phase ends and a new time period begins (in the play phase) when the adversary issues an oracle call with a time stamp $t+1$, where t is the current time.
- Let n be the total number of current honest users. The adversary is not allowed to issue an oracle call in the new time period, unless it is in the protocol phase and the last n oracle calls that it made where Activate oracle calls made to n different users and each of them returned the Done symbol.

We also add an additional restriction to the adversary: the adversary cannot issue a Play request on a honest user that was created in that period. The reason for this is that it is always possible for an adversary to do the following attack: at each time period, during the play phase, the adversary creates N users and, in the same phase, the adversary plays against them and defects.

2.4 Definition of Robustness

We may now define what it means for Π to be robust:

Definition 1 Π *is* robust *if $\mathcal{A}$ maximizes its utility by following Π; more formally, if the actions prescribed by Π form a subgame-perfect equilibrium*[7] *(cf. [7]).*

Additional desiderata. We view robustness as a necessary criterion for a trust-inference protocol to satisfy in order to be useful (if Π is not robust, than why would *any* party follow Π?). However, robustness alone is not enough. The following are some additional criteria that must be considered:

- The **expected utility** of Π is the utility a participant expects to achieve when everyone is honest. Clearly, higher expected utility is preferable.
- Π should ideally be **resilient to trembles** (see [6]) which occur when a player defects or fails to follow Π "by mistake", e.g., due to network faults rather than active cheating. The expected utility of Π may depend on the probability ε of trembles, and this should be taken into account.
- A protocol should also be **efficient at admitting new users**. Thus, even though new users may have to "pay their dues" [6, 9], the penalty for newcomers should not be so severe that it discourages users from joining altogether.
- Of course, the **efficiency** of Π (in terms of, say, the number of messages that must be sent) is also of interest.

As examples: a protocol that instructs all players to always defect is robust but has expected utility 0. The "grim trigger" strategy [7] (discussed below) is robust and achieves the best possible expected utility when $\varepsilon = 0$; however, it does not perform well when trembles occur with positive probability $\varepsilon > 0$. A protocol in which users do not interact with newcomers as long as reliable "veterans" are available may be robust but does not admit newcomers efficiently.

2.5 Extensions

The reader may well notice that the adversarial framework presented above is quite strong, and likely too pessimistic. Yet presenting such a strong framework has a number of advantages: (1) if a trust-inference protocol can be proven robust in such a strong model, it will certainly be robust in real-world adversarial environments; alternately, a "proof" that the model is too strong (in the sense that no reasonable and robust trust-inference protocols exist in that model) would be a very interesting and useful result; (2) the framework is general enough to encompass threats (such as coalitions, Sybil attacks, etc.) not typically modeled by previous work. Furthermore, (3) the framework is flexible enough to allow consideration of a number of more realistic threat models. We discuss some of these briefly now.

Network membership. In the model above, we have allowed the adversary to control the size of the network via NewUser calls. A more realistic model might assume that players continually join at some constant rate. The model may further assume that each party leaves the network with some probability α at each time period [6]. Note that using a

[7] Sometimes, we will relax this to require that it only form a Nash equilibrium.

model which assumes some constant turnover will automatically require a protocol to admit newcomers efficiently if it is to have high expected utility.

Network interactions. In the model above, we have allowed the adversary to control the trading partners of the honest parties via HonestPlay queries. While useful insofar as it models the worst-case behavior of the system, this clearly gives the adversary too much control. A more realistic model might have players paired off at random in a given time period. Furthermore, the model might assume that each player interacts exactly once during each time period; this would correspond to a *synchronous* network. To formalize the first restriction we say that the model is *random-pairing* if the HonestPlay oracle is replaced by a RandomPartner() oracle that returns the adversary's partner in this round. The adversary in round t can only call the Play oracle with the specified partner. To formalize the latter restriction (as an addition to the former), we say that the model is *synchronous random-pairing* if we require that the adversary calls the Play oracle exactly once for each time period.

No coalitions or Sybil attacks. Often, the simplifying assumption is made that the adversary acts alone (i.e., there are no coalitions) and can only act as a single player would (i.e., the adversary is not powerful enough to simulate the actions of multiple users). In general, we do not view such assumptions as realistic, although we agree that they simplify the analysis. In any case, it is easy to modify our model in the appropriate way (namely, by limiting the adversary to a single Play query per time period) to model this class of adversarial behavior. It is equally easy to modify our model so that a bound on the maximum coalition size is enforced. To formalize this, we say that the model is *single-identity attack* if the adversary is forced to use the same identity in all oracle calls made in the same time period.

3 Preliminary Results

We briefly sketch some preliminary results on the design of robust trust-inference protocols. These results reflect work in progress, and are important insofar as they demonstrate what is achievable in the model as sketched above, and also since (to the best of our knowledge) they are the first provably-robust protocols which do not assume any centralized infrastructure.

The first protocol we examine is the "grim trigger" strategy which mandates the following: all players cooperate until the first defection occurs. When defection occurs, the user who interacted with the defector in the previous round informs all players of this fact. Once a user hears that a defection has occurred, that user defects from then on. A more accurate description follows.

Protocol 1 (Grim Trigger)

A player that has never received a "grim trigger" message always cooperates.

If player i and j interact and i defects, then, in the following protocol phase, j sends a "grim trigger" message to everyone (inclding itself), either through a broadcast channel or point-to-point channels.

A player that has received a "grim trigger" message will always defect. It will also send a "grim trigger" message to everyone at every subsequent time period.

Lemma 1 *The "grim trigger" strategy is robust*[8]*, if the future discount factor δ is at least $\frac{1}{2}$, and it achieves optimal expected utility when the probability of trembles is 0, in the strongest adversarial model considered here.*

Proof We limit the proof to the two network models that we discussed in Section 2.2. Consider adversaries that are limited to N Play oracle calls per time period. Let δ be the discount factor.

Robustness We can design the following adversary G that is compliant with the protocol. In time period 0, G creates N honest users. Also, in each time period $t > 0$, G plays once with each of the honest users and it always cooperates; remember that G cannot play against those players at time 0. The total utility of G is:

$$u_G = \sum_{t=1}^{+\infty} N\delta^t = \frac{N\delta}{1-\delta} \qquad (1)$$

Now let's consider an arbitrary adversary A. We want to show that the utility u_A of A is at most u_G. There are two cases: either A never defects or there is a time period and a *Play* call in which it defects. If A never defects, then it can gain at most 1 unit of payoff at every game and therefore it can gain at most u_G. Therefore, let's assume that there is a time t' in which A defects in at least one of the games

[8] We note that its actions form a Nash equilibrium, not a subgame-perfect equilibrium.

and that A never defected before that. In any period $t < t'$ A can make at most N units of payoff (1 per game). For $t = t'$ A can make $2N$ units (2 per game). In the protocol phase of t', the honest players that have been defected on will send a message to all other players in the system. A can schedule the messages in the protocol phase, but it cannot prevent them to be delivered nor it can alter them. Therefore, in any period $t > t'$ all honest users will have received a "grim trigger" message and A will not make any utility at all. This means that the total utility of A can be at most:

$$u_A = \sum_{t=1}^{t'-1} N\delta^t + 2N\delta^{t'} = \frac{N\delta(1-\delta^{t'})}{1-\delta} + N\delta t' \quad (2)$$

Simple algebra shows that $u_A > u_G$ iff $\delta < \frac{1}{2}$.

Optimality If all the players follow the protocol, each player achieves an expected utility of 1 per interaction, which is the maximum possible. ■

We present this result only to indicate the feasibility of achieving robust solutions in our model. Of course, one problem with this strategy in practice is that it is not at all resilient to trembles.

Our second protocol is more interesting, and achieves a robust and efficient solution but still without any trusted third party. This protocol is a modification of the "pay-your-dues" (PYD) protocol of [6]. However, we stress that [6] assume a trusted authority who is also omniscient (and in particular knows the results of all interactions of the previous round), whereas we make no such assumption. Our adversarial model follows [6]: we assume synchronous trading, where in each round players are randomly paired. We also focus on single-player deviations, and assume that coalitions are not a concern (single-identity attack). We finally assume that the adversary has no control on the number of honest players, which is initially M_0 and the number of players in round i is $M_i = (1+\alpha)M_{i-1}$.

Our protocol Π is constructed as follows:

Protocol 2 (Broadcast PYD)

At the end of each round, each player broadcasts whether its partner from the previous round deviated or complied with the protocol.

A player i is defined to be a veteran *if a different player j broadcast a message stating that i was compliant in the previous round, no player broadcast a message stating that i was not compliant in the previous round and i broadcast* some *message about j in the previous round. All other players are called* newcomers *(note that this category includes both true newcomers as well as any players who deviated). If two different messages with the same sender identity are heard in the same round, all messages sent with that id in that round are ignored.*

In the following round, players trade as follows subject to the exceptions discussed below (this is similar to, but slightly different than [6]): if two veterans, they both cooperate; if two newcomers trade, they both defect; if a veteran trades with a newcomer, the veteran defects and the newcomer cooperates (the veteran's defection here is considered to be compliant with the protocol).

An exception to the above occurs if i is paired with j in two consecutive rounds. If this happens, then, for the purpose of their interaction, i is defined to be a veteran if it actually complied in such interaction in the previous round, otherwise i is defined to be a newcomer.

Another exception happens if i is paired with j in the current round, and in the previous round j broadcast a false complaint against i. In this case, i defects.

Note that we have essentially replaced the trusted party of [6] with a broadcast stage in which players announce whether their partner of the previous round deviated. However, *we take into account that players may* **lie** *when they broadcast this information* (in [6], the trusted party is assumed both to accurately know what really took place, as well as to reliably inform others of what occurred). In fact, the "exceptions" (above) is introduced exactly to ensure that lying will not increase the adversary's utility.

For simplicity, we analyze this protocol in a model with no trembles and we assume a broadcast channel.

Theorem 1 *Protocol 2 is a robust trust-inference protocol in the single-identity, synchronous random-pairing, broadcast channel model under the assumption that the number of players in the system is initially M_0 and that at round t is given by $M_t = (1+\alpha)M_{t-1}$, if the following condition holds:*

$$M_1 > 2 \quad \wedge \quad \delta \geq \frac{1}{2\left(1 - \frac{1}{M_1}\right)} \quad (3)$$

Sketch of Proof (Informal) Here we give an informal argument. A formal proof would require a more formal definition of the model and is given in the appendix.

We prove that the protocol is robust in the subgame-perfect equilibrium sense. Let G be any compliant adversary. More specifically, we prove that, for any history h and for any adversary A, starting from $t(h)$, A cannot achieve more utility than G.

We proceed in two steps. First we claim that it is sufficient to check the conditions for all adveraries A that may deviate from the protocol in round $t(h)$, but then comply in all future rounds. The proof of this is completely analogous to the One-Stage Deviation Condition Theorem (Theorem 4.2 in [7]) and we, therefore, omit the details. Then we show directly that a one-stage deviating adversary cannot do better than a compliant adversary.

Assume that h is a history and A is an adversary that complies with the protocol for all rounds $t > t(h)$. We want to compare the utility u_A achieved by A with the utility u_G achieved by G, starting at time $t(h)$. Since, in this model, the adversary has no control on the number of players, and given the particular structure of this protocol, the payoffs for A and G are the same in all rounds starting $t(h)+2$. Therefore, we need only to check that:

$$u_G(t(h)) + \delta u_G(t(h)+1) \geq u_A(t(h)) + \delta u_A(t(h)+1)$$

where $u_B(t)$ represents the expected payoff achieved by adversary B at round t and δ is the discount factor.

There are two cases: either (1) A is supposed to cooperate in round $t(h)$ and instead she defects, or (2) not.

Case (1) In round $t(h)$, A achieves one additional unit of payoff more than G. A's partner will broadcast a complaint against A, therefore A will be marked as a newcomer in the next round. (Even if A changes id from $i(t)$ to $i(t+1)$ and before changing her identity broadcasts a message that $i(t+1)$ was compliant, there will be no message broadcast by identity $i(t+1)$.)

In round $t(h)+1$, let p_N the probability that G's partner is a newcomer and let p'_N the probability that A's partner is a newcomer. Since both adversaries are compliant in this round:

$$\begin{aligned} u_G(t(h)+1) &= p_N \cdot 2 + (1-p_N) \cdot 1 = 1 + p_N \\ u_A(t(h)+1) &\leq p'_N \cdot 0 + (1-p'_N) \cdot (-1) = -1 + p'_N \\ u_G(t(h)+1) - u_A(t(h)+1) &\geq 2 + p_N - p'_N \end{aligned}$$

Remembering that the discount factor is δ, this implies that, for the equilibrium to hold, it must be that the gain in the current round $t(h)$ is no more than the discounted loss in the next round:

$$1 \leq \delta(2 + p_N - p'_N) \tag{4}$$

Given the assumption on the number of players in each round and the random pairing:

$$p_N = \frac{\alpha}{1+\alpha}$$

In the case of a broadcast channel, the best that the adversary can do during the protocol phase is to cause two honest players to become newcomers. She can do so by not broadcasting any message about her partner and by broadcasting a complaint message for another honest player. Therefore in the next round at most two of the honest players of the current round will be marked as newcomers. This means that:

$$p'_N \leq \frac{\alpha M_{t(h)} + 2}{(1+\alpha)M_{t(h)}} = p_N + \frac{2}{M_{t(h)+1}} \tag{5}$$

Plugging in this expression, we obtain the following equilibrium condition:

$$\delta \geq \frac{1}{2\left(1 - \frac{1}{M_{t(h)+1}}\right)} \tag{6}$$

Case (2) In this case, A achieves at most the utility that G achieves in round $t(h)$. The only way A can increase her own utility is using the protocol phase to cause some honest players at time $t(h)$ to become newcomers in the following round. A needs to be a veteran in the next round, in order to achieve more utility than G, therefore it cannot change identifier. If in the next round $t(h)+1$, A meets a honest player i that she caused to become a newcomer, then i will defect against A; otherwise A will obtain the same payoff as G would. This means that in this case A cannot do better than G.

Summing up In conclusion, equilibrium holds iff (6) is satisfied for all h. Since M_t is increasing, then equilibrium holds iff (3) holds. ■

A similar result can be obtained for the point-to-point channel model.

Theorem 2 *Protocol 2 is a robust trust-inference protocol in the single-identity, synchronous random-pairing, point-to-point channel model under the assumption that the number of players in the system is* *initially M_0 and that at round t is given by $M_t = (1+\alpha)M_{t-1}$, if the following condition holds:*

$$\delta \geq \frac{1+\alpha}{1+2\alpha} \tag{7}$$

Sketch of Proof (Informal) This proof proceeds analogously as the one of Theorem 1, except for the expression of p'_N in (5), Case (1). We can use the weaker bound $p'_N \leq 1$ and we thus obtain:

$$2 + p_N - p'_N \geq 2 + \frac{\alpha}{1+\alpha} - 1 = \frac{1+2\alpha}{1+\alpha}$$

Plugging in this expression in (4), we obtain (7), as needed. ■

4 Concluding Remarks

We stress that the framework presented here is a work in progress, and we do not claim that this is the final word on the subject. To the contrary, we hope that this paper inspires further work in this important area; that others will be motivated to refine and augment our model; and that researchers will attempt to design trust-inference protocols which can be rigorously proven to be robust within our framework. We feel strongly that the development and study of formal models for robust trust inference are necessary for future progress in this area.

Our work suggests a number of tantalizing open questions. First, can robust trust-inference protocols with very low communication requirements (in particular, not requiring broadcast) be designed? Alternately, can one show the impossibility of designing very efficient yet robust protocols in a particular adversarial environment? We conjecture that efficient and robust trust inference is impossible (when no trusted authority is assumed) within adversarial models which allow arbitrary-size coalitions/Sybil attacks. It would be wonderful to formalize and rigorously prove (or disprove) this conjecture within the framework given here.

References

[1] K. Aberer and Z. Despotovic. Managing trust in a peer-2-peer information system. In *Proc. Intl. Conf. on Information and Knowledge Management*, 2001.

[2] C. Buragohain, D. Agrawal, and S. Suri. A game theoretic framework for incentives in p2p systems. In *Proc. 3rd Intl. Conf. on Peer-to-Peer Computing*, 2003.

[3] E. Damiani, S.D.C. di Vimercati, S. Paraboschi, P. Samarati, and F. Violante. A reputation-based approach for choosing reliable resources in peer-to-peer networks. In *9th ACM Conf. on Computer and Communications Security*, 2002.

[4] J.R. Douceur. The sybil attack. In *1st Intl. Workshop on Peer-to-Peer Systems*, 2002.

[5] First interdisciplinary symposium on online reputation mechanisms, Apr 2003. See http://www.si.umich.edu/˜presnick/reputation/symposium.

[6] E. Friedman and P. Resnick. The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, 1998.

[7] D. Fudenberg and J. Tirole. *Game Theory.* MIT Press, 2002.

[8] S. Goldwasser, S. Micali, and R. Rivest. A digital signature scheme secure against adaptive chosen-message attacks. *SIAM J. Computing*, 17(2):281–308, 1988.

[9] S.D. Kamvar, M.T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proc. 12th Intl. Conf. on the World Wide Web*, 2003.

[10] K. Lai, M. Feldman, I. Stoica, and J. Chuang. Incentives for cooperation in peer-to-peer networks. In *Workshop on Economics of Peer-to-Peer Systems*, 2003.

[11] Seungjoon Lee, Rob Sherwood, and Bobby Bhattacharjee. Cooperative peer groups in NICE. In *IEEE Infocom*, 2003.

[12] R. Lethin. Reputation. In Oram [15], chapter 17.

[13] R. Levien and A. Aiken. Attack-resistant trust metrics for public key certification. In *USENIX Security Symposium*, 1998.

[14] P. Nixon and S. Terzis, editors. *Proc. 1st Intl. Conf. on Trust Mgmt.* Springer-Verlag, 2003. LNCS, vol. 2692.

[15] A. Oram, editor. *Peer-to-Peer: Harnessing the Power of Disruptive Technologies.* O'Reilly, 2001.

[16] Project NICE at the University of Maryland. http://www.cs.umd.edu/projects/nice/.

[17] M.K. Reiter and S. Stubblebine. Authentication metric analysis and design. *ACM Trans. Info. and System Security*, 2(2):138–158, 1999.

[18] Reputations research network. See http://databases.si.umich.edu/reputations/.

[19] P. Resnick, K. Kuwabara, B. Zeckhauser, and E. Friedman. Reputation systems. *Comm. ACM*, 43(12):45–48, 2000.

[20] M. Waldman, L.F. Cranor, and A. Rubin. Trust. In Oram [15], chapter 15.

A A formal model for Theorem 1

In this section, we give a formal description for the model assumed in Theorem 1. We need to define the notions of *history*, *protocol*, *adversary* and *experiment*. Then we can define what it means for a protocol to be robust and present a more precise proof of the theorem.

A.1 A model for the protocol

We will give an ad-hoc definition of *history*, as opposed to use a standard game-theoretical definition. This will make our exposition easier.

Definition 2 (History) *A history is a tuple $h = (h^0, \ldots, h^{t-1})$, where, for each $\tau = 0, \ldots, t-1$, h^τ is a set of tuples (i, p_i, a_i, m_i) where i is a user in the system at time τ, p_i is the partner of user i at time τ, a_i is the action of user i at time τ and m_i is the message broadcast by user i at time τ. For any user id i, the local history of i extracted from h (denoted by $L_i(h)$) is:*

$$L_i(h) = (L_i(h^0), \ldots, L_i(h^\tau)) \tag{8}$$

where

$$L_i(h^\tau) = (i, p_i, a_i, a_{p_i}, (m_j)^{\forall j}) \tag{9}$$

Definition 3 (Protocol) *In this model, a protocol is a PPT algorithm Π, that can be run*

- *either with three inputs: a user id i, a local history h_i and a second user id p_i and in this case it returns an action:*

$$a_i \leftarrow \Pi(i, h_i, p_i) \tag{10}$$

- *or with four inputs: the same inputs as above and an action a' and, in this case, it returns a message:*

$$m_i \leftarrow \Pi(i, h_i, p_i, a') \tag{11}$$

The way the protocol is used is the following. For every user i and in every round t, let h_i be the local history of i up to that point. The user i gets paired at random with another user p_i to play a prisoner's dilemma instance. If the user follows Π, then the user runs $\Pi(i, h_i, p_i)$ and this will return the action a_i taken by i. Then the user learns the opponent move a' and as a consequence it broadcasts a message m_i obtained as $\Pi(i, h_i, p_i, a')$.

We now introduce an adversary, which is a single non-colluding player, which can change its identity over time and may deviate arbitrarily from the protocol. In order to do that, we need to augment the local history of the adversary with the vector $\xi = (\xi^0, \ldots, \xi^{t-1})$, representing the actual identities that the adversary has controlled over time. We say that ξ is identity sequence compatible with h if for each pair of rounds (τ, τ'), if $\tau \neq \tau'$, then either $\xi^\tau = \xi^{\tau'}$ or $\xi^{\tau'}$ does not appear in h^τ (this means that, if the adversary uses a certain identity at some time τ', then no other player uses that identity).

Definition 4 (Adversary) *An adversary is any PPT A, with the following structure. It can be ran:*

- *either with two inputs:*

$$i \leftarrow A(\xi, h) \tag{12}$$

  in which case it takes the past history and returns the identity that the adversary wants to take in this round.

- *or with three inputs:*

$$a \leftarrow A(\xi, h, p) \tag{13}$$

  in which the adversary learns its partner and chooses its action

- *or with four inputs:*

$$m \leftarrow A(\xi, h, p, a') \tag{14}$$

  in which the adversary learns the partner's action and produces a message.

In this model, we want to define what it means to run an experiment, given a protocol Π, an adversary A, starting from a setting where previous history is h and continuing for Δ rounds.

Definition 5 (Experiment) *Let h be a history, Δ be a non-negative integer, ξ be an identity sequence compatible with h, A be an adversary and Π a protocol. The experiment $\mathsf{Exp}_{A,\Pi}(\xi, h, \Delta)$ is defined as follows. For each $\tau = t, \dots, t + \Delta - 1$ do the following.*

- *Let $M^{\tau-1} = |h^{\tau-1}|$ the number of players at the time t. Then $M^\tau = \lceil (1+\alpha)M^{\tau-1} \rceil$. Let $H^{\tau-1}$ the set of the identities of honest players at time $\tau - 1$, i.e. the set of all i such that some $(i, \dots) \in h^{\tau-1}$ and $\xi^{\tau-1} \neq i$. Then H^τ is $H^{\tau-1}$ incremented with $M_\tau - M_{\tau-1}$ arbitrarily chosen identities not appearing in $(h^0, \dots, h^{\tau-1})$.*
- *Execute*
$$\xi^\tau \leftarrow A((\xi^0, \dots, \xi^{\tau-1}), (h^0, \dots, h^{\tau-1}))$$
(new adversary identity).
- *Let π be a uniformly random pairing of elements in $H^\tau \cup \{\xi^\tau\}$.*
- *Execute*
$$a^\tau_{\xi^\tau} \leftarrow A((\xi^0, \dots, \xi^{\tau-1}), (h^0, \dots, h^{\tau-1}), \pi(\xi^\tau)).$$
- *For each $i \in H^\tau$, let $k^\tau_i = L_i(h^0, \dots, h^{\tau-1})$ and execute:*
$$a^\tau_i \leftarrow \Pi(i, k^\tau_i, \pi(i))$$
- *Execute*
$$m^\tau_{\xi^\tau} \leftarrow A((\xi^0, \dots, \xi^{\tau-1}), (h^0, \dots, h^{\tau-1}), \pi(\xi^\tau), a^\tau_{\pi(\xi^\tau)}).$$
- *For each $i \in H^\tau$, execute:*
$$m^\tau_i \leftarrow \Pi(i, k^\tau_i, \pi(i), a^\tau_{\pi(i)})$$
- *Let*
$$h^\tau = \{(i, \pi(i), a^\tau_i, m^\tau_i) : i \in H^\tau \cup \{\xi^\tau\}\}$$

The output of the experiment is:
$$(\xi', h') = ((\xi^0, \dots, \xi^{t+\Delta-1}), (h^0, \dots, h^{t+\Delta-1})$$

Finally, the experiment $\mathsf{Exp}_{A,\Pi}(\xi, h)$ is the version of $\mathsf{Exp}_{A,\Pi}(\xi, h, \Delta)$ with $\Delta = \infty$.

We will say that an adversary is admissable if, for all ξ and h, $A(\xi, h)$ produces an identity that is *not* in $H^{t(h)}$.

Definition 6 (Compliant adversary) *The compliant adversary G with protocol Π is the adversary defined as follows.*

- *It never changes identity ($G(\xi, h) = \xi^{t(h)-1}$).*
- *It follows the protocol*
$$G(\xi, h, p) = \Pi(\xi^{t(h)-1}, L_{\xi^{t(h)-1}}(h), p)$$
$$G(\xi, h, p, a') = \Pi(\xi^{t(h)-1}, L_{\xi^{t(h)-1}}(h), p, a')$$

Similarly, one can define what it means for an adversary A to be compliant in some round t: the two conditions above hold only for any h such that $t(h) = t$.

Definition 7 (Utility of the adversary)
$$U(\xi, h) = \sum_{\tau=0}^{t} \delta^\tau u(a^\tau_{\xi^\tau}, a^\tau_{p_{\xi^\tau}}) \tag{15}$$
where $u(a, a')$ is the payoff of a player who plays action a in a PD game with another player who plays action a' (entry of the payoff matrix).

A.2 Subgame Perfection

We now extend the definition of *subgame-perfect equilibrium* to our model. The classical definition of subgame perfection [7, Definition 3.5] is, unfortunately, not sufficient for our purposes, because it is restricted to multi-round games where all players learn the entire game history up to the previous round. In contrast, in our model a honest user i learns only a *local* portion $L_i(h)$ of the past history h. Instead of extending the framework of multi-stage games to a more general set of games that include our model, we decide to give an ad-hoc definition of subgame perfection for our purposes.

Definition 8 (Subgame perfection) *A protocol Π is called subgame perfect if, for any adversary A, for any history h, for any identity vectory ξ compatible with h, it is:*
$$\mathrm{E}[U(\mathsf{Exp}_{A,\Pi}(\xi, h))] \leq \mathrm{E}[U(\mathsf{Exp}_{G,\Pi}(\xi, h))] \tag{16}$$
where G is the compliant adversary.

We can now extend the classical One-Stage Deviation Condition Theorem [7, Theorem 4.2] to our notion of subgame perfection.

Theorem 3 *Protocol Π is subgame perfect if and only if, for any history h, for any identity vectory ξ compatible with h, for any adversary A' that may deviate at round $t = t(h)$ but is compliant in any following round, it is:*

$$\mathrm{E}[U(\mathsf{Exp}_{A',\Pi}(\xi, h))] \leq \mathrm{E}[U(\mathsf{Exp}_{G,\Pi}(\xi, h))] \qquad (17)$$

where G is the compliant adversary.

Sketch of Proof (Informal) The proof of this theorem is completely analogous to the proof of the analogous Theorem 4.2 in [7]. The "only if" portion of the proof is immediate.

For the "if" part, we show that if Π is not subgame perfect, then we can construct h', ξ' and an adversary A' that deviates only in round $t = t(h')$, such that A' does better than the compliant adversary. Towards this goal, assume that h and ξ are as in the theorem statement and A is an adversary which deviates from the protocol (possibly in all rounds) and achieves expected utility *strictly greater* than G.

I can show that there exists a time T such that the adversary B, which behaves like A up to round T and then complies afterwards, achieves utility greater than G. This follows from the fact that contributions to the utility in the far future count very little, as long as $\delta < 1$.

Then I can construct a sequence of adversaries $B^T, B^{T-1}, \ldots, B^{t-1}$ such that B^τ is the same as A for all rounds up to τ, and complies afterwords. Note that $B^T = B$. It must be the case that, for at least one of the values of τ, B^τ does better than $B^{\tau-1}$. For that value of τ, set h' to be the history generated by running $B^{\tau-1}$ starting from h up to round $\tau - 1$, let ξ' be the corresponding identity vector and let A' be B^τ. This completes the proof.

■

B Proof of Theorem 1.

Proof We have to show that the protocol Π is subgame perfect. Using Theorem 3, we state that it is sufficient to check that for any history h, for any identity vector ξ and for any adversary A that complies at rounds $t > t(h)$, the expected utility $\mathrm{E}[U(\mathsf{Exp}_{A,\Pi}(\xi, h))]$ of A is no more than the corresponding expected utility of the compliant adversary G.

In order to prove this, consider an execution $\mathcal{E}$ of the experiment $\mathsf{Exp}_{A,\Pi}(\xi, h)$. Many different executions are possible, depending on the choice of the random pairings and the random coins of the adversary and honest players. Fixing an execution $\mathcal{E}$ means fixing a value for all those random choices. Consider also the execution $\mathcal{E}'$ of $\mathsf{Exp}_{G,\Pi}(\xi, h)$, such that the pairings and the random coins for all players are the same in $\mathcal{E}$ and $\mathcal{E}'$.

I distinguish two cases for $\mathcal{E}$: either (1) A is supposed to cooperate in round $t(h)$ and instead she defects, or (2) not. Formally, let $\xi^{t(h)} = A(\xi, h)$ be the identity chosen by A at round $t(h)$ in $\mathcal{E}$; let $p^{t(h)}$ A's partner at time $t(h)$. Case (1) happens if:

$$\Pi(\xi^{t(h)}, L_{\xi^{t-1}}(h), p^{t(h)}) = C$$
$$A(\xi, h, p^{t(h)}) = D$$

where C and D are the actions "cooperate" and "defect"; case (2) happens otherwise.

Let $U_A(t), U_G(t)$ be the (not discounted) payoffs achieved by A and G during an execution of $\mathcal{E}$ and $\mathcal{E}'$ respectively, at round t. Since, in this model, the adversary has no control on the number of players, and given the particular structure of this protocol, the payoffs for A and G are the same in all rounds starting $t(h) + 2$. We will show that $U_A(t) + \delta U_A(t + 1) \leq U_G(t) + \delta U_G(t + 1)$, in expectation conditioned on case (1) and for any $\mathcal{E}$ satifying case (2).

Case (1). In round $t(h)$, A achieves 1 more payoff than G. A's partner will broadcast a complaint against A, therefore A will be marked as a newcomer in the next round. (Even if A changes id from ξ^t to ξ^{t+1} and before changing her identity broadcasts a message that ξ^{t+1} was compliant, there will be no message broadcast by identity ξ^{t+1}.)

In round $t(h) + 1$, let p_N the probability that G's partner is a newcomer and let p'_N the probability that A's partner is a newcomer, under random choices of $\mathcal{E}$ ($\mathcal{E}'$), conditioned on case (1) happening. Since both adversaries are compliant in this round:

$$\mathrm{E}[U_G(t(h) + 1)] = p_N \cdot 2 + (1 - p_N) \cdot 1 = 1 + p_N$$
$$\mathrm{E}[U_A(t(h) + 1)] \leq p'_N \cdot 0 + (1 - p'_N) \cdot (-1) = -1 + p'_N$$
$$\mathrm{E}[U_G(t(h) + 1)] - \mathrm{E}[U_A(t(h) + 1)] \geq 2 + p_N - p'_N$$

where we used the fact that the random choice on the partner at time $t(h) + 1$ is independent of the random choices on which A and Π decisions depend at round $t(h)$.

Given the assumption on the number of players in each round, the random pairing and the fact that, if

G is the adversary, all players in the game at time $t(h)$ will become veterans in the next round:

$$p_N = \frac{\alpha}{1+\alpha}$$

In the case of a broadcast channel, the best that the adversary can do during the protocol phase is to cause two honest players existing at round $t(h)$ to become newcomers. She can do so by not broadcasting any message about her partner and by broadcasting a complaint message for another honest player. Therefore in the next round at most two of the honest players of the current round will be marked as newcomers. This means that:

$$p'_N \leq \frac{\alpha M_{t(h)} + 2}{(1+\alpha)M_{t(h)}} = p_N + \frac{2}{M_{t(h)+1}}$$

Plugging in this expression, we obtain the following condition:

$$\delta \geq \frac{1}{2\left(1 - \frac{1}{M_{t(h)+1}}\right)}$$

which is sufficient to make the expected utility of A to be no more than the one of G, conditioned on case (1).

Case (2) In this case, A achieves at most the utility that G achieves in round $t(h)$. The only way A can increase her own utility is using the protocol phase to cause some honest players at time $t(h)$ to become newcomers in the following round. A needs to be a veteran in the next round, in order to achieve more utility than G, therefore it cannot change identifier. If in the next round $t(h)+1$, A meets a honest player i that she caused to become a newcomer, then i will defect against A; otherwise A will obtain the same payoff as G would. This means that in this case A cannot do better than G.

Summing up In conclusion, equilibrium holds iff (6) is satisfied for all h. Since M_t is increasing, then equilibrium holds iff (3) holds.

■