

# Sybilproof Transitive Trust Protocols

Paul Resnick  
School of Information, University of Michigan  
Ann Arbor, MI 48109 USA  
presnick@umich.edu

Rahul Sami  
School of Information, University of Michigan  
Ann Arbor, MI 48109 USA  
rsami@umich.edu

## ABSTRACT

We study protocols to enable one user (the *principal*) to make potentially profitable but risky interactions with another user (the *agent*), in the absence of direct trust between the two parties. In such situations, it is possible to enable the interaction indirectly through a chain of credit or “trust” links. We introduce a model that provides insight into many disparate applications, including open currency systems, network trust aggregation systems, and manipulation-resistant recommender systems. Each party maintains a trust account for each other party. When a principal’s trust balance for an agent is high enough to cover potential losses from a bad interaction, direct trust is sufficient to enable the interaction. Allowing indirect trust opens up more interaction opportunities, but also expands the strategy space of an attacker seeking to exploit the community for its own ends. We show that with indirect trust exchange protocols, some friction is unavoidable: any protocol that satisfies a natural strategic safety property that we call *sum-sybilproofness* can sometimes lead to a *reduction* in expected overall trust balances even on interactions that are profitable in expectation. Thus, for long-term growth of trust accounts, which are assets enabling risky but valuable interactions, it may be necessary to limit the use of indirect trust. We present the hedged-transitive protocol and show that it achieves the optimal rate of expected growth in trust accounts, among all protocols satisfying the sum-sybilproofness condition.

## Categories and Subject Descriptors

K.4.4 [Computing Milieux]: Computers and Society—*Electronic Commerce*

## General Terms

Theory, Economics

©ACM, 2009. This is the author’s version of the work. It is posted here by permission of the ACM for your personal use. Not for redistribution. The definitive version appears in the Proceedings of the ACM EC09Conference, Palo Alto, USA, July 2009. DOI: <http://doi.acm.org/10.1145/1566374.1566423>  
EC’09, July 6–10, 2009, Stanford, California, USA.  
Copyright 2009 ACM 978-1-60558-458-4/09/07.

## Keywords

sybilproof, transitive trust, indirect reciprocity, open currency, reputation system, recommender system

## 1. INTRODUCTION

Potentially valuable transactions often require one or both parties to take a risk. A new employee may not work out or the employer may go bankrupt. An online book seller may take a payment but not ship the goods. A restaurant recommendation from an apparently disinterested stranger on the street may turn out to be from the restaurant owner’s cousin. When the transacting parties do not have reason to trust each other directly, the benefits of such transactions may be missed: the employee may not be hired; the buyer may not buy the book from the untrusted seller; potentially valuable restaurant recommendations may be ignored.

Sometimes, intermediaries can provide an indirect chain of trust between parties who do not have sufficient direct trust to risk a transaction. For example, previous employers or personal references can vouch for the skill and integrity of a potential employee. Amazon.com may place its own reputation at stake by allowing a used book seller to list on its site. In online systems ranging from cryptographic key authentication [17] to recommender systems [3] to peer-to-peer networks [6], the idea of a web of trust has emerged, possibly involving longer chains than a single intermediary. That is, principal  $p$  may decide to trust agent  $a$  because  $p$  trusts  $c$ , who trusts and vouches for  $b$ , who trusts and vouches for  $a$ .

Two natural questions arise in such systems of transitive trust. First, under what conditions should the principal  $p$  trust agent  $a$ , given a set of certifications from other parties? For example, in cryptographic key authentication, even if  $p$  is pretty sure that a particular key belongs to  $c$ , it’s not clear whether  $p$  should trust that some other key belongs to  $b$  simply because  $c$  says so. Second, if  $p$  undertakes a risky transaction with  $a$ , based in part on certifications from  $b$  and  $c$ , how should the willingness to provide or accept future certifications change when the current transaction turns out well or badly?

With respect to the first question, there is a growing literature on aggregating trust statements into a metric that is robust against collusive reporting by a number of identities [1, 10, 6]. For example, the Advogato trust metric [10] is designed to certify a user’s expertise in software development, on a scale ranging from “Novice” to “Master”, based on the network of users vouching for other users’ skill. One design feature of this scheme is that, regardless of the num-

ber of collusive, malicious identities, at any point in time the total damage to the system (measured in terms of inappropriate certifications) is limited to the total trust mistakenly given by honest programmers to the attack identities.

This paper addresses the second question, of how the various parties should adjust, after the current transaction concludes, their willingness to trust each other in the future. One intuition is that if a transaction backed by a chain of trust goes badly, distrust will propagate throughout the chain. For example, if a professor provides a glowing reference letter for a student, who is then hired and performs poorly, reference letters from that professor will have less credibility in the future, and the professor will also be less willing to provide such a glowing letter for that student.

Another intuition is that when a transaction backed by a chain of trust goes well, there may not be a symmetric but opposite propagation of increased trust. In particular, in a setting where attackers can create many sybil identities [2], there is a danger that the principal and agent may be controlled by the same attacker: they may produce an artificial report of a positive transaction in order to influence one or more of the intermediaries to increase their trust of other entities. For example, suppose a stranger asks for your opinion about a restaurant, then later reports that the restaurant was good. It may be natural, psychologically, to treat this as confirmation of the quality of the restaurant or the good taste of the stranger, but there is no logical reason to do either: Doing so would put you at risk if the stranger was in fact a con artist trying to induce your trust.

We offer a formal model and prove theorems based on the two insights above. The model is of interactions with one-sided risk: the principal  $p$  may gain or lose value depending on the hidden action of an agent  $a$ . Each entity in the system maintains a trust account for each other entity. A protocol decides which transactions should be undertaken, with each transaction specifying an interaction and a post-interaction trust updating function. We show that if a protocol’s allowable transactions all satisfy a *sum-sybilproofness* property and the protocol never allows any trust account to fall below 0, the damage from an attacker will be limited. We show that a seemingly natural protocol that passes  $p$ ’s declared profit through intermediaries to  $a$  is not sum-sybilproof. We also present an alternative, the *hedged-transitive* protocol, that is sum-sybilproof.

The hedged-transitive protocol is asymmetric in its effects, making more decrements to trust accounts after a negative outcome than the increments it makes after successful transactions. We show further that any sum-sybilproof protocol must be similarly asymmetric. The most striking consequence is that the expected sum of total trust account balances among all parties will decline if a transaction is sufficiently risky, even if it has positive expected profit for  $p$ . If we think of the balances in the trust accounts as assets that enable the community to undertake risky individual interactions while still limiting the damage that attackers can do, there is no way to enable some risky but profitable (in expectation) interactions without depleting the communal trust assets (in expectation).

This simple formal model provides insights into a number of applications. We illustrate this by elaborating its application to manipulation-resistant recommender systems, online trade environments with two-sided risk, and open currency/IOU systems.

**Robust recommenders:** The model and results apply directly to a scenario of using transitive trust in a recommender system, while preventing manipulation. In recent work [13], we developed a recommender system that is provably robust against attacks by an attacker who can create fake identities and inject spurious ratings, perhaps to boost or suppress selected items. The *Influence Limiter* algorithm works by tracking a reputation for each pair  $p, a$  of users; this reputation limits how much influence  $a$  can have on  $p$ ’s next recommendation. The reputation is updated as  $a$ ’s information turns out to be helpful or hurtful to recommendations of items to  $p$ .

The drawback of this approach is that, until  $a$  has built up a reputation with  $p$ , much of  $a$ ’s information is lost due to the influence limits. One approach to reducing the information loss is to think of these reputations as trust accounts, and allow chaining of trust through intermediaries. The results in this paper show that the most natural way of using transitivity can lead to more sophisticated attacks on the recommender system over time. The hedged-transitive protocol provides a safe alternative to exploit transitive influence, but is not a panacea because it may reduce the net value of trust accounts when transactions are risky.

**Two-sided risk:** The model can also be applied to a scenario where transactions involve two-sided risks: both sides have something to gain if the transaction goes well and something to lose if it does not. Such transactions neatly decompose into two independent, one-sided transactions: the overall transaction takes place only if each side, taking on the role of principal, has enough trust, either direct or mediated by others, to proceed; as each side declares an outcome independently it triggers an updating of trust accounts. It is even possible for the two sides of a transaction to be separated in time. For example, consider a customer buying a book from an Internet retailer. The first one-sided interaction consists of the retailer as principal receiving a payment from the buyer as agent. (That interaction entails little risk to the seller so little trust is needed.) The second one-sided interaction has the buyer as principal, getting value if the seller as agent sends the correct book in a timely fashion, but otherwise not.

**Open currency systems:** The model also provides insight into open currency systems such as Yootles [12] and Ripplepay [15]. Such systems allow anyone to create an account, and issue IOUs to pay for goods. They build on the fact that users are willing to extend some amount of credit to others they know. The goal of these sites is to facilitate profitable interactions (trade, or otherwise) without requiring the “double-coincidence” of barter systems, that both parties simultaneously have something of equal value to exchange [7]. They attempt to do so without the overhead of actual monetary exchange and in barter settings where people may have services but not money to pay each other. For example, in a neighborhood, people might mow lawns, cook meals, or do small house repairs for others.

The trust accounts and records of current transactions can be viewed as a form of “memory” [8], which substitutes for money as an enabler of indirect exchanges. If Amy has to pay Peter for mowing her lawn, Amy can issue an IOU, denominated in some numeraire currency such as dollars. The IOU is a promise to offer a good or perform some service, at a future time, of value equal to the amount of the IOU.

If Peter had previously issued an IOU to Amy, then instead of Amy issuing an IOU, she could simply cancel Peter’s previous IOU to her.

More generally, these systems are intended to allow indirect barter. Whenever there is a cycle of people who have issued IOUs, each one to the next person in the cycle, they can all agree to cancel the IOUs. Each person cancels one obligation owed to him or her and has one of his or her obligations canceled.

Each user maintains a credit limit for every other user, and will not accept IOUs if doing so would exceed that user’s credit limit. This maps neatly to our model:  $p$ ’s trust balance for  $a$  would be  $p$ ’s credit limit for  $a$ , minus any outstanding IOUs  $a$  has issued to  $p$ , plus any outstanding IOUs  $p$  has issued to  $a$ . Thus, when  $p$  cancels one of  $a$ ’s IOUs, he increases his trust balance for  $a$  by that amount. A traditional monetary exchange can be viewed as a special case of our model, in which there is a distinguished intermediary (the Central Bank) who commands infinite trust from all other users. In this setting, the trust “assigned by the Bank” to a user  $a$  models  $a$ ’s current money holdings.

Without a Central Bank, our model suggests how credit could be extended in chains of trust: when Amy does not have sufficient credit with Peter, she can issue an IOU to Bob and Bob can issue one to Peter. Under our hedged-transitive protocol, the IOUs should be accepted only if Bob’s trust balance for Amy and Peter’s trust balance for Bob are larger than the amount of the IOUs. Immediately, Peter will decrease Bob’s trust balance and Bob will decrease Amy’s by the amount of the IOU. The chaining procedure ensures the following safety property: An attacker cannot extract more value by creating a set of sybil identities, so long as honest participants do not place any direct trust in the new sybil identities [11].

The chaining or “rippling” procedure addresses how to safely use mediated credit links for a given configuration of credit limits and current IOU holdings, but it does not specify any norms for how these credit limits are updated following a transaction. The results of this paper address that problem. If, as a result of the interaction, Amy receives something of value from Peter (say lawn-mowing services), then she may increase her trust accounts by up to the amount of the value to her of the service (presumably more than the amount of the IOU she just issued.) In particular, Amy should increase her trust balance for Peter by at least the amount of the IOU, and can also increase her trust in Bob by some amount, so that she is willing to provide some service to Bob in the future, as a reward for his serving as a trust intermediary in this transaction. Note, however, that if Amy later fails to provide any return services to anyone to redeem the IOU, there will be an outstanding IOU from Bob to Peter and thus some of Peter’s credit limit to Bob will be tied up and the community will be less able to solve the double-coincidence problem.

*Other Related Work.* Notions of sybilproofness for mechanisms have been studied in the context of auctions as well, where the term *false-nameproof* has been used [16]. Although the attack form is similar to ours in that an attacker creates and controls a number of identities, there are two significant distinctions: we study bilateral interactions (perhaps mediated by others agents), and we analyze the effect of sybilproofness on future trust availability.

The fact that easy creation of pseudonyms, and strategic protection against attack by agents with multiple pseudonyms, can lead to losses of efficiency has been earlier noted in the contexts of reputation systems [4] and recommender systems [14]. There is a key difference between our results in this paper and those earlier results: The earlier results captured the loss of efficiency that arises when agents or mechanisms have to be skeptical of unproven users they are interacting with. This paper focuses specifically on the inefficiency that is necessary to resist manipulation in indirect trust mechanisms. Such mechanisms attempt to enable interactions, through indirect trust, where there is insufficient direct trust between the principal and agent. Paradoxically, in order to avoid sybil attacks, we find that, for sufficiently risky transactions (but still with positive expected value), it is necessary to destroy more trust than is created in expectation. Thus, to avoid sybil attacks, using indirect trust may impair the ability to carry out further interactions, even those that would have been supported through direct trust, had that trust not been squandered on vouching for parties in indirect trust transactions that had bad outcomes.

In a recent paper, Landa *et al.* [9] have independently studied sybilproof use of trust networks, in the context of peer-to-peer systems. Although primarily addressing the question of enabling a principal to trust an agent based on chains of trust, they also show that trust can potentially be depleted at a faster rate through indirect interactions. They suggest a heuristic to ameliorate this, which is similar to the cyclic exchange transaction we describe in Remark 1.

*Structure of this paper.* In section 2, we introduce the abstract model of transitive trust exchange protocols, define the sum-sybilproofness and no negative holdings conditions, and show that together they imply limited damage from attackers. In section 3, we present instances of protocols that satisfy, or do not satisfy, this strategic property, as well as some transformations that preserve the sum-sybilproofness property. Section 4 contains our main result, and discusses its ramifications; several generalizations and extensions of this result are presented in section 5. Finally, in section 6, we conclude and outline directions for future research.

## 2. THE BASIC MODEL

There is a set of people and a set of users (more accurately, usernames)  $U$ . The set  $U$  is common knowledge, but the people behind them are unknown. The people are either *honest members* who each control a single username, or *the attacker*, who controls the remaining set of usernames. Each person knows the username(s) he or she controls, but no one else knows this linkage, *i.e.*, there is perfect pseudonymity. We assume that the cost of creation of a pseudonym is small.

The community activity happens through a series of interactions. Each interaction has two associated usernames: the principal  $p$  and the agent  $a$ . For simplicity, we will refer to the choices as if they were made by  $p$  and  $a$ , and payoffs as if they accrue to them, though strictly speaking the actions are made by the people controlling  $p$  and  $a$  and the payoffs accrue to the people as well. If the parties choose to interact, the agent  $a$  chooses a hidden action that affects the probability that the interaction will be successful (that is, that  $p$  will perceive the outcome as good).

If  $p$  is honest, the payoff is +1 for a successful interaction, and −1 for an unsuccessful interaction. Agent  $a$  receives no

payoff (though there may be a payoff from a corresponding one-sided transaction where  $a$  acts as the principal). The attacker is assumed to be adversarial to the community, so that his payoff is the negative of the sum of all honest members' payoffs. If the principal  $p$  is one of the attacker's usernames, there is no payoff to any agent.

The users have access to trust accounts and ledgers to record transactions. Each user maintains his or her own trust account for each other user. Formally, we let  $R_{uv}$  denote the amount of trust that  $u$  assigns to  $v$ . We use  $\mathbf{R} = \{R_{uv}\}$  to denote the set of all trust account balances at a given time.

Formally, we define a transaction to include not only the interaction but also a plan for updating trust accounts for the principal, agent, and any intermediaries, after the principal announces a transaction outcome.

**DEFINITION 1.** A transaction  $T$  is a tuple  $(p, a, S, \Delta^+, \Delta^-)$ , where

- $a \in U$  is the agent,  $p \in U$  is the principal, and  $S \subseteq U, a, p \in S$  is the set of users involved in this transaction, including intermediaries.
- At some point in the future,  $p$  declares the interaction outcome to be  $+$  (successful) or  $-$  (unsuccessful). The matrices  $\Delta^+ = \{\Delta_{uv}^+ | u, v \in S\}$  and  $\Delta^- = \{\Delta_{uv}^- | u, v \in S\}$  define the adjustments made to the trust accounts of all parties, including intermediaries: If the outcome is  $+$ , each  $R_{uv}$  has  $\Delta_{uv}^+$  added to it, and if the outcome is  $-$ , each  $R_{uv}$  has  $\Delta_{uv}^-$  added to it. Note that the entries in  $\Delta^+$  and  $\Delta^-$  can be positive, negative, or zero.

We will assume that transactions are serialized, so the outcome of one is known before the next one is initiated. We consider the effects of concurrency in Section 5.

We equate one unit of trust with one unit of payoff or loss and define direct transactions as follows:

**DEFINITION 2.** A **direct transaction** is a transaction  $T_{pa} = (p, a, S, \Delta^+, \Delta^-)$  with  $S = \{p, a\}$ , and  $\Delta_{pa}^+ = 1, \Delta_{ap}^+ = 0, \Delta_{pa}^- = -1, \Delta_{ap}^- = 0$ .

We are interested in identifying classes of safe transactions for the settings we consider. In other words, we envisage the community adopting a protocol on what sets  $T$  of transactions are permissible in a given configuration.

**DEFINITION 3.** A **network trust exchange protocol**  $\mathcal{P} : \mathbf{R} \mapsto \{T\}$  specifies, for each trust configuration  $\mathbf{R}$ , a set of allowable transactions  $\mathcal{T}_{\mathcal{P}}(\mathbf{R})$ . The protocol is assumed to include all feasible direct transactions, i.e., the set of allowed transactions  $\mathcal{T}_{\mathcal{P}}(\mathbf{R})$  must include all direct transactions  $T_{pa}$  such that  $R_{pa} \geq 1$ .

**DEFINITION 4.** A network trust exchange protocol  $\mathcal{P}$  satisfies the **no negative holdings** property if there is no sequence of allowed transactions, starting from a state in which all trust balances are non-negative, that results in a trust balance  $R_{uv}$  being negative. Formally, for any  $\mathbf{R}$  such that  $\forall u, v, R_{uv} \geq 0$ , and  $T \in \mathcal{T}_{\mathcal{P}}(\mathbf{R})$ , we must have:

$$\forall u, v \Delta_{uv}^+ \leq R_{uv} \text{ and } \Delta_{uv}^- \leq R_{uv}$$

Intuitively, the no negative holdings property allows us to think of trust account balances as limits on potential future losses. In particular, the initial account balance  $R_{pa}$  should reflect the maximum loss that  $p$  is willing to sustain with  $a$ .

## 2.1 Sum-sybilproofness

Transactions may involve a number of usernames, and the participants in a transaction can never be certain which usernames belong to the attacker. A natural design goal is therefore to identify a class of transactions that in some way limit the damage the attacker can cause to the honest agents. In this section, we motivate and define our formal specification of this design goal.

For any single transaction, it is inherently possible that it winds up unsuccessful even if all participating agents are honest. In the case of direct transactions, the trust accounts provided a way for a user to limit the damage by the attacker. Purely through direct transactions, an attacker who controls a set  $Y$  of usernames can inflict a total damage of  $\sum_{v \in Y, u \in U \setminus Y} R_{uv}$ . We would like the more general protocols we develop to not expand the attacker's power.

The key strategic property we require is that the network exchange protocol be *sybilproof*: an attacker should not be able to gain utility (i.e., inflict additional damage on the honest community) by participating in a sequence of transactions allowed by the protocol. As direct transactions are always permitted, the trust balance at any point reflect the total potential net damage the attacker can cause going forward. Thus, we need to ensure that an allowed transaction does not unduly increase the total trust assigned to members of  $Y$ , regardless of which set  $Y$  is actually controlled by the attacker. Formally, we define:

**DEFINITION 5.** A transaction  $T = (p, a, S, \Delta^+, \Delta^-)$ , which specifies a two-party interaction and functions for updating trust balances after successful and unsuccessful transactions, satisfies the **sum-sybilproofness** property if, for every possible subset  $H \subset S$  of honest users, and each possible outcome with value  $x$  declared by the principal, the following conditions hold:

$$\begin{aligned} \text{if } p \notin H : & \quad \sum_{u \in H, v \in \bar{H}} \Delta_{uv}^x \leq 0 \\ \text{if } p \in H : & \quad \sum_{u \in H, v \in \bar{H}} \Delta_{uv}^x \leq x \end{aligned}$$

For the specific  $\pm 1$  outcomes of the transactions we consider in this paper, this can be expanded as:

$$\begin{aligned} \text{if } p \notin H : & \quad \sum_{u \in H, v \in \bar{H}} \Delta_{uv}^+ \leq 0 \\ \text{if } p \in H : & \quad \sum_{u \in H, v \in \bar{H}} \Delta_{uv}^+ \leq 1 \\ \text{if } p \notin H : & \quad \sum_{u \in H, v \in \bar{H}} \Delta_{uv}^- \leq 0 \\ \text{if } p \in H : & \quad \sum_{u \in H, v \in \bar{H}} \Delta_{uv}^- \leq -1, \end{aligned}$$

where  $\bar{H} = S \setminus H$  is the complement of  $H$ .

We say that a protocol  $\mathcal{P}$  is a **sum-sybilproof protocol** if every transaction  $T$  permitted by the protocol is sum-sybilproof.

This property states that, for any set  $H$  of honest users, the additional trust issued by members of  $H$  is no more than the profit or loss incurred by members of  $H$ , if any. Note that the definition does not include trust assigned by

non-members of  $H$  at all. The motivation for this is that trust assigned by an attacker identity need not be honored in future, and thus may not be worth anything.

We use the term sum-sybilproofness to distinguish this from the related concepts of value-sybilproofness and rank-sybilproofness defined by Cheng and Friedman [1]; the latter concepts constrain the maximum rank or value any single attack identity can obtain, whereas the *total* value (of all trust assigned to attackers) is relevant in our context.

Note that sum-sybilproofness only governs the *change* in trust balances following a transaction; it does not specify how the initial trust balances were determined. If  $u$  assigns trust to  $v$ , she is essentially issuing credit to  $v$ . In a community in which agents cannot be linked to their usernames, it is natural to ask where this initial seeding of credit comes from. There are several possible reasons that fit with the spirit of our model: There might be a cost to creating the username  $v$ , and each honest agent might assign  $v$  trust equivalent to a fraction of that cost. This cost might be explicitly imposed on a new identity, perhaps through a “money-burning” mechanism [5]. Alternatively, even in the absence of a cost,  $u$  might assign a very small amount of trust to each unknown username; this allows  $u$  to profit from interactions with  $v$  in the long run, while limiting her regret if the interactions are not profitable. In this paper, we are agnostic to the source of the initial allocations. However, the fact that they are likely to be small gives added importance to our study of the rate of growth of the trust balances:  $v$  and  $u$  cannot engage in a sizeable interaction until their mutual trust balances are sufficient to cover their respective potential losses.

## 2.2 Computational Infrastructure

In this subsection, we briefly discuss the computational implications of our model.

We assume that participants can communicate with each other, perhaps through an online forum they are all logged into. They can also record transactions in a secure ledger: at the start of the transaction, all participants approve it, and at the end, the principal enters her declaration and the ledger is updated.

In practice, this ledger service can be distributed, so that all trust assigned by username  $u$  is maintained by  $u$ . This does not make the system significantly more vulnerable: the honest people would run their ledgers honestly, and never use others’ ledgers to make their trust decisions.

One reasonable way to implement such ledgers would be to use public-key encryption for users to sign transactions, so that the attacker cannot inject transactions on behalf of usernames it does not control. Note that the requirements are significantly lighter than a standard public-key infrastructure because the system is not required to register which user controls a given username. A person who wishes to create a new username can simply generate a new random public-key pair, and use the public key itself as the username. This eliminates the certification step from the public key distribution. Of particular importance, it eliminates the need to have a trusted root certification authority. Indeed, if it were feasible to deploy a single certification authority trusted by everyone, through which every agent certified their primary username, a true sybil attack could be prevented (although similar collusive attacks could still take place). In many communities, however, there may not be a

universally trusted root authority, making this certification process infeasible.

We abstract away the details of the underlying communication layer. If the communication layer is controlled by the attacker, they may be able to carry out denial-of-service attacks on the honest community, but not force exploitative transactions on them. On the other hand, if the communication layer is controlled by a trusted party, it may be possible in principle to build a certification process into this lower layer that makes attacks difficult, for example, by using geographic or network address as a component of the username. Our goal is to analyze transactions in the absence of any linkage information between usernames and their controlling persons, so we do not include any such assumptions.

## 3. SAFE TRUST EXCHANGE PROTOCOLS

In this section, we present some network interaction protocols that are sum-sybilproof.

### 3.1 Direct interaction

The simplest class we consider is the class of all direct, unmediated, interaction transactions. Formally, we define the direct interaction protocol  $\mathcal{P}_{dir}$  as the protocol that, given a configuration  $\mathbf{R}$ , restricts allowable transactions to only the direct transactions  $T_{pa}$  where  $R_{pa} \geq 1$ . In words: an allowable transaction credits the agent’s account with one unit of the principal’s trust if successful, or debits the agent’s account by one if unsuccessful. The condition  $R_{pa} \geq 1$  ensures that the trust balance is never negative after a transaction.

Transactions in the direct interaction protocol are clearly sum-sybilproof, because the only trust balance that is updated is  $R_{pa}$ , and thus the only relevant condition is when  $H = \{p\}$ . The change in  $R_{pa}$  reflects the payoff that  $p$  derives from this transaction, so the constraint is satisfied for  $H = \{p\}$ . Thus, a community that restricted itself to direct transactions would be robust against attacks of the form we consider.

### 3.2 Transitive protocols

The drawback of the direct protocol  $\mathcal{P}_{dir}$  is that the class of allowed transactions is too restrictive. For example, if  $v$  has information or services of value to  $u$ , but  $u$  has not assigned trust directly to  $v$ , the value will never be realized. Thus, a protocol that allows transaction mediated by other usernames can lead to greater efficiency.

#### 3.2.1 A symmetric protocol

The most natural extension that one might consider is to allow transitive transactions according to the protocol  $\mathcal{P}_2$ , described as follows. For a given  $\mathbf{R}$ ,  $\mathcal{P}_2$  includes all transactions of the form  $T_{pwa}$  such that  $R_{wa} \geq 1, R_{pw} \geq 1$ . The set of usernames involved in  $T_{pwa}$  consists of the agent  $a$ , the principal  $p$ , and an intermediary  $w$ . The trust updates  $\Delta^+$  and  $\Delta^-$  are defined as follows:

- If the outcome is  $+$ ,  $R_{pw}$  is incremented by 1 and  $R_{wa}$  is incremented by 1.
- If the outcome is  $-$ ,  $R_{pw}$  is decremented by 1 and  $R_{wa}$  is decremented by 1.
- In either case, all other trust balances are left unchanged.

Informally, we can think of  $p$  as risking a transaction with  $a$  because of transitive trust:  $p$  trusts  $w$  and  $w$  trusts  $a$ . Similarly, we can intuitively think of  $\mathcal{P}_2$  updating the trust accounts by passing  $p$ 's declared profit or loss through  $w$  to  $a$ . After a good transaction,  $p$  trusts  $w$  more and  $w$  trusts  $a$  more; after a bad transaction, they trust less.

Somewhat surprisingly, the protocol  $\mathcal{P}_2$  is *not* sum-sybilproof:

LEMMA 1. *The protocol  $\mathcal{P}_2$  is not sum-sybilproof.*

PROOF. Consider the set  $H = \{w\}$ , i.e.,  $a$  and  $p$  are controlled by an attacker. Then, if the transaction is declared successful by  $p$ , we have  $R_{wa}$  incremented by 1, even though  $w$  has not experienced a payoff, thus violating the sum-sybilproofness constraint for  $H = \{w\}$ . Intuitively, if  $a$  and  $p$  are controlled by the attacker,  $p$ 's reported profit in the transaction with  $a$  will trick  $w$  into increasing her trust account for  $a$ , who could then exploit that later to suggest one more harmful action to  $w$  than  $a$  would otherwise have been able to. The fake transaction will increase  $p$ 's trust account for  $w$ , but that is of no value since  $p$  is an attacker and can simply disappear after the fake transaction.  $\square$

### 3.2.2 Hedged-transitive protocols

We present another protocol, the hedged-transitive protocol  $\mathcal{P}_H$ , that plugs the loophole in  $\mathcal{P}_2$  by *reducing*  $w$ 's trust of  $p$  after  $p$  declares a successful transaction. For a given  $\mathbf{R}$ ,  $\mathcal{P}_H$  includes all transactions of the form  $T'_{pwa}$  such that  $R_{wa} \geq 1$ ,  $R_{pw} \geq 1$ , and  $R_{wp} \geq 1$ . The set of usernames involved in  $T'_{pwa}$  consists of the agent  $a$ , the principal  $p$ , and an intermediary  $w$ . The trust updates  $\Delta^+$  and  $\Delta^-$  are defined as follows:

- If the outcome is  $+$ ,  $R_{pw}$  is incremented by 1,  $R_{wa}$  is incremented by 1, and  $R_{wp}$  is *decremented* by 1.
- If the outcome is  $-$ ,  $R_{pw}$  is decremented by 1 and  $R_{wa}$  is decremented by 1. ( $R_{wp}$  is left unchanged.)
- In either case, all other trust balances are left unchanged.

Note that the hedged transaction  $T'_{pwa}$  is asymmetric in that the trust updates on a  $-$  outcome are not the exact opposites of the corresponding changes on a  $+$  outcome, even though the interaction payoff is symmetric. In Section 4, we will show that this asymmetry is unavoidable if we have intermediated sum-sybilproof transactions.

THEOREM 2. *The hedged-transitive protocol  $\mathcal{P}_H$  is sum-sybilproof.*

PROOF. Intuitively, the transaction  $T'_{pwa}$  eliminates the strategic hazard in  $T_{pwa}$ : If  $a$  and  $p$  are attack usernames, and  $p$  declares the interaction successful, then  $R_{wp}$  is decremented to compensate for  $R_{wa}$  being incremented, so the attacker does not have a net increase in assigned trust. Formally, we can consider six cases, one for each non-empty subset  $H$  of  $S$ , and verify that the trust balances have not been inappropriately raised to violate the constraint:

- *Case (i)  $H = \{a\}$ :* Trivial, because assignments of  $a$ -trust do not change.
- *Case (ii)  $H = \{w\}$ :* As discussed above,  $R_{wa} + R_{wp}$  does not increase.

- *Case (iii)  $H = \{p\}$ :*  $R_{pa}$  is unchanged, and  $R_{pw}$  is increased only when  $p$  obtained a payoff of  $+1$ , and decreased otherwise.
- *Case (iv)  $H = \{a, w\}$ :*  $R_{wp}$  is never increased, only potentially decreased.
- *Case (v)  $H = \{w, p\}$ :*  $R_{pa}$  is unchanged, and  $R_{wa}$  is increased only when  $p$  obtained a payoff of  $+1$ , and decreased otherwise.
- *Case (vi)  $H = \{p, a\}$ :*  $R_{aw}$  is unchanged.  $R_{pw}$  is increased only when  $p$  obtained a payoff of  $+1$ , and decreased otherwise.

$\square$

$\mathcal{P}_H$  specifies that the principal put all of the surplus gained from a positive outcome into the trust account for the intermediary. The sum-sybilproof result shows that an attacker can never do more net damage than the initial trust balances honest people placed in the attacker's sybils. The attacker, however, can potentially claim all of the surplus from any successful transactions where a sybil merely acts as the trust intermediary. That is, each such successful transaction adds one to the trust account of a sybil, who can then later use it to cause a bad outcome of a direct transaction between the principal and the sybil.

An obvious countermeasure is not to pass all the surplus of successful transactions to the trust accounts of intermediaries. We generalize from the simple hedged-transitive protocol  $\mathcal{P}_H$  to a family of protocols. The transaction  $T_{pwa}(\alpha, \beta)$  is identical to the hedged transaction  $T'_{pwa}$  when the outcome is negative. When the outcome is positive, however, the additional trust distributions are different. The principal distributes additional trust of  $\alpha \leq 1$  instead of 1. This additional trust is apportioned between the intermediary  $w$  and the agent  $a$ , with  $w$  getting a fraction  $\beta$ . The intermediary passes on any additional trust it receives to its trust account for  $a$ , but correspondingly decrements its trust for  $p$ .

DEFINITION 6. *For any  $\alpha \in [0, 1]$  and any  $\beta \in [0, 1]$ , the generalized hedged-transitive transaction  $T_{pwa}(\alpha, \beta)$  is a transaction with principal  $p$ , agent  $a$ , and intermediary  $w$ . The trust updates are defined as follows:*

- *If the outcome is  $+$ ,  $R_{pw}$  is incremented by  $\alpha\beta$ ,  $R_{pa}$  is incremented by  $\alpha(1 - \beta)$ ,  $R_{wa}$  is incremented by  $\alpha\beta$ , and  $R_{wp}$  is decremented by  $\alpha\beta$ .*
- *If the outcome is  $-$ ,  $R_{pw}$  is decremented by 1,  $R_{wa}$  is decremented by 1, and all other trust balances are left unchanged.*

*The generalized hedged-transitive protocol  $\mathcal{P}_H(\alpha, \beta)$  is a protocol that allows all direct transactions  $T_{pa}$  with  $R_{pa} \geq 1$ , and all generalized hedged-transitive transactions  $T_{pwa}(\alpha, \beta)$  with  $R_{pw} \geq 1$ ,  $R_{wa} \geq 1$ , and  $R_{wp} \geq \alpha\beta$ .*

In order to prove that the generalized hedged-transitive protocol is sum-sybilproof, we first present some basic transformations that allow us to derive new sum-sybilproof transaction forms from known ones. These transformations will also be useful in the proof of our main negative result.

LEMMA 3. *The following operations preserve sum-sybilproofness:*

- (i) dominance: *If  $T = (p, a, S, \Delta^+, \Delta^-)$  is sum-sybilproof, and  $T' = (p, a, S, \tilde{\Delta}^+, \tilde{\Delta}^-)$  is such that, for all  $u, v \in S$ ,  $\tilde{\Delta}_{uv}^+ \leq \Delta_{uv}^+$  and  $\tilde{\Delta}_{uv}^- \leq \Delta_{uv}^-$ , then  $T'$  is also sum-sybilproof.*
- (ii) crossover: *If  $T = (p, a, S, \Delta^+, \Delta^-)$  and  $T' = (p, a, S, \tilde{\Delta}^+, \tilde{\Delta}^-)$  are two sum-sybilproof transactions, then the transaction  $T'' = (p, a, S, \Delta^+, \tilde{\Delta}^-)$  is also sum-sybilproof.*
- (iii) mixing: *If  $T = (p, a, S, \Delta^+, \Delta^-)$  and  $T' = (p, a, S, \tilde{\Delta}^+, \tilde{\Delta}^-)$  are two sum-sybilproof transactions, then the transaction  $T'' = (p, a, S, \lambda\Delta^+ + (1-\lambda)\tilde{\Delta}^+, \lambda\Delta^- + (1-\lambda)\tilde{\Delta}^-)$ , for any  $0 \leq \lambda \leq 1$  is also sum-sybilproof.*
- (iv) permutation: *Suppose  $T = (p, a, S, \Delta^+, \Delta^-)$  is sum-sybilproof. Let  $\pi$  be any permutation of  $S$  that leaves  $a$  and  $p$  fixed; we denote by  $\pi(i) \in S$  the username that  $i$  is mapped to under this permutation. Then, define  $\tilde{\Delta}^+$  by  $\tilde{\Delta}_{\pi(i)\pi(j)}^+ \stackrel{\text{def}}{=} \Delta_{ij}^+, \forall i, j \in S$ , and  $\tilde{\Delta}^-$  likewise. The transaction  $T_\pi = (p, a, S, \tilde{\Delta}^+, \tilde{\Delta}^-)$  is also sum-sybilproof.*

PROOF.

- (i) *dominance:* All the constraints in the definition of sum-sybilproofness place upper bounds on sums of  $\Delta^+$  and  $\Delta^-$  terms. Thus, replacing  $\Delta_{uv}^+$  (or  $\Delta_{uv}^-$ ) with a smaller value preserves sum-sybilproofness.
- (ii) *crossover:* Observe that the constraints with outcome  $+$  do not affect  $\Delta^-$ , and the constraints with outcome  $-$  do not affect  $\Delta^+$ .
- (iii) *mixing:* All sum-sybilproofness constraints involve linear functions of the variables in  $\{\Delta_{uv}^+\}$  and  $\{\Delta_{uv}^-\}$ . Thus, if they are satisfied by  $\Delta^+$  and  $\tilde{\Delta}^+$  (viewed as vectors in  $\mathfrak{R}^{|S|(|S|-1)}$ ) they are also true for convex combinations of  $\Delta^+$  and  $\tilde{\Delta}^+$ .
- (iv) *permutation:* Permuting the names of the mediators  $S \setminus \{p, a\}$  merely corresponds to permuting the constraints in the sum-sybilproofness condition, and thus preserves the property.  $\square$

COROLLARY 4. *The generalized hedged-transitive protocol is sum-sybilproof.*

PROOF. We can get to the transaction  $T_H(\alpha, \beta)$ , starting from the hedged-transitive transaction  $T'_{pwa}$  and the direct transaction  $T_{pa}$ , through a sequence of transformations that, by Lemma 3, preserve sum-sybilproofness. First, we use the mixing property to construct a  $(\alpha\beta, 1 - \alpha\beta)$  mix of  $T'_{pwa}$  and  $T_{pa}$ . This would yield  $\Delta_{pw}^+ = \Delta_{wa}^+ = \alpha\beta$ ,  $\Delta_{wp}^+ = -\alpha\beta$ , and  $\Delta_{pa}^+ = 1 - \alpha\beta$ . Next, the dominance property allows us to replace  $\Delta_{pa}^+$  by the smaller quantity  $\alpha - \alpha\beta = \alpha(1 - \beta)$ . Finally, we use the crossover property to couple the resulting  $\Delta^+$  for successful transactions with the update function  $\Delta^-$  from  $T'_{pwa}$  from for unsuccessful transactions.  $\square$

## 4. IMPOSSIBILITY RESULT

In section 3, we presented a class of mediated transactions, the hedged-transitive transactions, that satisfy the sum-sybilproofness property. One feature of these transactions is their asymmetry. In particular, transaction  $T'_{pwa}$  decrements  $R_{wp}$  if successful, but does not increment it if unsuccessful. Apart from being unnatural, this decrement has a detrimental effect on communities without attackers: it leads to the trust assignments (*i.e.*, available credit) growing slower than they would in the basic transitive protocol  $T_{pwa}$ . In this section, we show that this slower growth of total trust is unavoidable if we want to satisfy the basic strategic robustness captured in sum-sybilproofness.

THEOREM 5. *Let  $T = (p, a, S, \Delta^+, \Delta^-)$  be a transaction that satisfies sum-sybilproofness. Then, if  $\Delta_{pa}^- \geq 0$  and  $\Delta_{ap}^- \geq 0$ , both the following must hold:*

- (i)  $\sum_{v, u \in S, v \neq u} \Delta_{uv}^+ \leq 1$
- (ii)  $\sum_{v, u \in S, v \neq u} \Delta_{uv}^- \leq -2$

Note that, if  $R_{pa}$  and  $R_{ap}$  are initially 0 (the minimum permissible level), negative  $\Delta_{pa}^-$  and  $\Delta_{ap}^-$  are not allowed, and thus, the conditions of the theorem are satisfied. Other situations where  $p$  and  $a$  have some direct trust but that trust is not put at risk as the basis for the transaction also satisfy the conditions of the theorem.

PROOF. (i) Intuitively, we must protect each  $u \neq p$  from the possibility that the other usernames are all controlled by the attacker. Then, we must insure that  $u$  does not issue additional trust when the attacker reports a positive transaction outcome. Thus, for any  $u \neq p$ ,  $\sum_{v \neq u} \Delta_{uv}^+ \leq 0$  by the sum-sybilproofness constraint with  $H = \{u\}$ . When  $u = p$ , and the real transaction has a positive outcome,  $p$  can issue additional trust, but not more than the utility gain of 1; thus,  $\sum_{v \neq p} \Delta_{pv}^+ \leq 1$  in order to satisfy the sum-sybilproofness constraint with  $H = \{p\}$ . Simply summing over all  $u$ , we get that  $\sum_u \sum_{v \neq u} \Delta_{uv}^+ \leq 1$ , which is the desired result.

(ii) The second part is slightly more involved. We first observe that if there are no intermediaries ( $S = \{p, a\}$ ), there is no sum-sybilproof transaction in which  $\Delta_{pv}^- \geq 0$ . Next, consider the case in which there is only one mediating node  $w$ , and so  $S = \{a, w, p\}$ . Informally, the intuition we would like to use is as follows:  $p$  does not know if  $w$  is colluding with  $a$ , so it has to reduce  $w$ 's trust balance by 1. Similarly,  $p$  and  $w$  together cannot be exploited by  $a$ , so  $w$  has to reduce  $a$ 's trust balance by 1.

Formally, we apply the sum-sybilproofness conditions to get the following inequalities:

$$\begin{aligned} \Delta_{pw}^- + \Delta_{pa}^- &\leq -1 \quad (H = \{p\}) \\ \Delta_{wp}^- + \Delta_{ap}^- &\leq 0 \quad (H = \{w, a\}) \\ \Delta_{wa}^- + \Delta_{pa}^- &\leq -1 \quad (H = \{p, w\}) \\ \Delta_{aw}^- + \Delta_{ap}^- &\leq 0 \quad (H = \{a\}) \end{aligned}$$

Adding them together, we get:

$$\begin{aligned} \left[ \sum_{v,u,v \neq u} \Delta_{uv}^- \right] + \Delta_{pa}^- + \Delta_{ap}^- &\leq -2 \\ \left[ \sum_{v,u,v \neq u} \Delta_{uv}^- \right] &\leq -2 - \Delta_{pa}^- - \Delta_{ap}^- \leq -2 \end{aligned}$$

The same intuition works for the case with two or more mediating nodes, but the analysis is slightly more complex because there may be trust updates between mediating nodes. In order to simplify the analysis, we first show that we can restrict attention to transactions in which the mediators are all symmetric:

*Claim:* If there is a sum-sybilproof transaction  $T$  with  $S = \{p, a, w_1, w_2, \dots, w_r\}$  for which  $\sum_{v,u} \Delta_{uv}^- = m$ , there is another sum-sybilproof transaction  $T'$  with the same set in which all nodes  $w_i$  are perfectly symmetric, *i.e.*,  $T'$  is invariant under permutation of the intermediate nodes.

*Proof of claim:* Consider all permutations  $\pi$  of  $S$  that leave  $a$  and  $p$  fixed. By the permute property of Lemma 3, the transaction  $T_\pi$  is sum-sybilproof; it also has the same sum  $\sum \Delta_{uv}^- = m$ . Next, repeatedly using the mixture property in Lemma 3, we take the average  $T'$  of all  $r!$  transactions  $T_\pi$ , preserving both sum-sybilproofness and the sum  $m$ .  $T'$  is clearly symmetric across the usernames  $w_1, \dots, w_r$ , as every permutation of these usernames have equal weight in the average.

Using this claim, we can restrict attention to transactions  $T = \{p, a, S = \{p, a, w_1, \dots, w_r\}, \Delta^+, \Delta^-\}$  with the following properties: For all  $w_i, w_j$ , we have:  $\Delta_{pw_i}^- = \alpha$ ;  $\Delta_{w_i p}^- = \alpha'$ ;  $\Delta_{aw_i}^- = \beta$ ;  $\Delta_{w_i a}^- = \beta'$ ;  $\Delta_{w_i w_j}^- = \gamma$ , for some parameters  $\alpha, \alpha', \beta, \beta', \gamma$ .

If  $\gamma \leq 0$ , we must have  $m \leq -2$ , because we can coalesce all the intermediaries into a single username  $w$  while not reducing the sum  $\sum_{v,u} \Delta_{uv}^-$ . The resulting transaction would satisfy sum-sybilproofness, as coalescing a set of usernames is equivalent to relaxing the sum-sybilproofness condition, requiring it to hold only for sets  $H$  that contain all usernames  $w_i$  or no usernames  $w_i$ . We would thus be left with a sum-sybilproof transaction with one intermediary, for which we have shown  $m \leq -2$ . Thus, we assume that  $\gamma > 0$  in the remainder of the proof.

Suppose that  $r$  is even,  $r = 2k$ . Then, we have

$$m = 2k(\alpha + \alpha' + \beta + \beta') + 2k(2k - 1)\gamma + \Delta_{pa}^- + \Delta_{ap}^-$$

Consider the set  $G = \{p, w_1, \dots, w_k\} \subset S$ . From sum-sybilproofness, we have the following relations:

$$\sum_{v \in \overline{G}, u \in G} \Delta_{uv}^- = k(\alpha + \beta') + k^2\gamma + \Delta_{pa}^- \leq -1 \quad (H = G)$$

$$\sum_{v \in G, u \in \overline{G}} \Delta_{uv}^- = k(\alpha' + \beta) + k^2\gamma + \Delta_{ap}^- \leq 0 \quad (H = \overline{G})$$

Adding together and multiplying by 2, we have:

$$\begin{aligned} &2k(\alpha + \alpha' + \beta + \beta') + 4k^2\gamma + 2(\Delta_{pa}^- + \Delta_{ap}^-) \\ &= m + 2k\gamma + \Delta_{pa}^- + \Delta_{ap}^- \leq -2 \\ &m \leq -2 - 2k\gamma - \Delta_{pa}^- - \Delta_{ap}^- < -2 \end{aligned}$$

The case in which  $r = 2k + 1$  is odd can be shown similarly, except that we need to consider two sets:  $G_1$  with  $k$  intermediate nodes, and  $G_2$  with  $k + 1$  intermediate nodes.  $\square$

*Remark 1:* It is obvious that the condition  $\Delta_{pa}^- \geq 0$  is required in the statement of Theorem 5, since if  $p$  could reduce his trust in  $a$  directly (without creating a negative balance), a direct trust interaction would be possible. Rather surprisingly, the condition  $\Delta_{ap}^- \geq 0$  is also required in some cases: If  $a$  can reduce  $p$ 's trust by 1, it is possible to find a transaction that is sum-sybilproof such that  $\sum \Delta_{uv}^- = -1$ .

One way to construct such a transaction is to compose a three-way trust cycle reversal with a direct transaction. If  $a$  trusts  $p$ ,  $p$  trusts  $u$ , and  $u$  trusts  $a$ , we can reverse the direction of the cycle of trust, without changing anyone's total vulnerability to attackers. That is,  $a$  increases her trust of  $u$  and reduces her trust of  $p$ ;  $u$  increases her trust of  $p$  and reduces her trust of  $a$ ;  $p$  increases his trust of  $a$  and reduces his trust of  $u$ . Following this,  $p$ 's trust balance for  $a$  is sufficient to carry out a direct transaction. If it is unsuccessful,  $p$  will decrement the trust for  $a$  he just incremented in the trust cycle reversal. That would leave two increments from the cycle reversal ( $a$ 's trust of  $u$  and  $u$ 's trust of  $p$ ) as well as the three decrements. Thus, one feasible combined transaction, from the initial state, would have:

$$\Delta_{au}^- = 1; \Delta_{up}^- = 1; \Delta_{ap}^- = -1; \Delta_{ua}^- = -1; \Delta_{pu}^- = -1$$

Independently, Landa *et al.* [9] proposed a similar idea of reducing indirect transactions to direct transactions by rotating cycles of trust where possible.

*Consequences of Theorem 5:* Consider situations in which two agents with usernames  $a$  and  $p$  do not assign each other trust, but can potentially profit from an interaction. If they do so with a transaction that is not sum-sybilproof, an attacker could set up interactions that are damaging to the honest agents as a group. On the other hand, if the community norms specify that only sum-sybilproof transactions are permitted, Theorem 5 shows that there is an unavoidable asymmetry: if the interaction creates a surplus value of 1, at most one unit of the asset (total trust) can be created, whereas if the interaction leads to a loss of 1, at least 2 units of the asset must be removed from the system.

The consequences of this asymmetry are most striking when there is a modest bias towards profitable interactions. For example, suppose that the outcome of an interaction between honest  $a$  and  $p$  is generated by a random process with probability  $p = 0.6$  of success. Then, the expected value created in the interaction is  $0.6 - 0.4 = 0.2$ , but the expected increase in trust as a result of this interaction is at most  $0.6 - 0.8 = -0.2$ , *i.e.*, there is an expected *decrease* in the total trust. This is not merely a nominal problem: If the assignment  $R_{uv}$  is reduced, it implies that  $u$  has a lower credit limit for  $v$  in future transactions, which may eliminate a potentially profitable interaction in the future. (If, instead, the value of all trust accounts is adjusted to compensate, so that a smaller holding leads to the same effective credit, an attacker would be able to do an increasing amount of damage over time just by preserving his holdings.) Thus, a long sequence of transactions that decrease the trust balances (in expectation) can lead to slower growth of value in the community.

Note that this is a first-order analysis that looks only at total trust assignment (available credit), but not at how the trust is distributed among the community. The distribution will have an effect on long-run efficiency, but this effect

will be highly dependent on how profitable interactions are themselves distributed. The analysis of total trust still suggests that there are situations in which any indirect trust exchange protocol can be detrimental in the long run.

## 5. EXTENSIONS TO THE BASIC MODEL

For clarity, we have presented our main result in a stylized model in which all interactions have a payoff of  $\pm 1$ , the outcome is completely unobservable, and the agent does not incur a cost. The result easily generalizes along all of these dimensions, as long as the instance includes the uncertainties of the stylized model to some extent.

*Asymmetric bets and multiple outcomes.* Suppose that the value to the principal is  $+x$  when the outcome is successful, and  $-y$  otherwise. Then, the following result follows immediately from Theorem 5, as the constraints on  $\Delta^+$  and  $\Delta^-$  are linear in  $x$  and  $y$  respectively:

**COROLLARY 6.** *Suppose that a risky transaction  $T = (p, a, S, \Delta^+, \Delta^-)$  will result in the principal receiving a payoff of  $+x$  if successful or  $-y$  if unsuccessful. Then, if  $\Delta_{pa}^- \geq 0$  and  $\Delta_{ap}^- \geq 0$ , we must have:*

$$\sum_{v,u \in S, v \neq u} \Delta_{uv}^+ \leq x ; \quad \sum_{v,u \in S, v \neq u} \Delta_{uv}^- \leq -2y \quad \square$$

Similarly, we can extend the definition of sum-sybilproofness to transactions with more than two possible outcomes. As the constraints on each outcome are independent, it follows from Theorem 5 that the total net increase in trust following an outcome with positive value  $x$  is no more than  $x$ ; for an outcome with negative value  $-y$ , there is a net reduction in total trust balances of at least  $2y$ .

*Partially direct transactions.* Theorem 5 applies in the setting in which  $a$  and  $p$  had no credit relationship, *i.e.*,  $R_{pa} = R_{ap} = 0$ . The result immediately extends to the case in which  $R_{pa}, R_{ap}$  combined are smaller than the transaction value:

**COROLLARY 7.** *Suppose that a risky transaction  $T = (p, a, S, \Delta^+, \Delta^-)$  will result in the principal receiving a payoff of  $+1$  if successful or  $-1$  if unsuccessful. Then, if  $\Delta_{pa}^- + \Delta_{ap}^- \geq -c$ , for  $c < 1$ , we must have:*

$$\sum_{v,u \in S, v \neq u} \Delta_{uv}^+ \leq 1 ; \quad \sum_{v,u \in S, v \neq u} \Delta_{uv}^- \leq c - 2$$

**PROOF.** Follows directly by putting  $\Delta_{pa}^+ + \Delta_{pa}^- \geq -c$  in the proof of Theorem 5.  $\square$

*Verifiable costs and benefits.* Often, the agent has an observable cost  $c$ . Further, some aspects of the outcome may be public information rather than subjectively revealed through the principal. We can model the outcome as having a public value  $x$  in addition to the uncertain value  $\pm 1$ . We can decouple the certain parts of this transaction, transforming it into the combination of two transactions (1) A non-risky transaction  $T_1$  that includes trust assignment based on the commonly known cost and value, and (2) a risky interaction without observable cost or benefits. The latter component will still be a source of friction, provided that there are outcomes in which the subjective component results in negative value.

*Concurrent transactions.* For simplicity, we assumed that transactions are serialized. This is a restriction on the power of the attacker, and thus, the impossibility result in Theorem 5 would clearly hold even if we relaxed this restriction by admitting the possibility of concurrent transactions. Our positive result that the hedged-transitive protocol satisfying the sum-sybilproof property can be extended to the concurrent setting with conservative timing of updates: At the start of a transaction, each trust balance has to be updated (or encumbered) to cover the worst-case reduction in the balance that could arise from that transaction; they may later be increased when the actual outcome is revealed.

*Longer chains of transitivity.* Theorem 5 can be strengthened for the case in which there is neither a direct relationship between  $a$  and  $p$ , nor a path of length 2. In this case, safely executing a transaction will require a longer chain of connections between  $a$  and  $p$ .

**DEFINITION 7.** *A transaction  $T = (p, a, S, \Delta^+, \Delta^-)$  is an  $\ell$ -chain if the set  $S$  can be written as  $S = \{w_0 \stackrel{def}{=} p, w_1, w_2, \dots, w_{\ell-1}, w_\ell \stackrel{def}{=} a\}$  such that*

$$\Delta_{uv}^- \leq 0 \Rightarrow \exists i \text{ s.t. } v = w_i, u \in \{w_{i-1}, w_{i+1}\}$$

We can now prove a stronger version of Theorem 5 for this special case:

**THEOREM 8.** *Let  $T = (p, a, S, \Delta^+, \Delta^-)$  be a transaction that is an  $\ell$ -chain, and satisfies sum-sybilproofness. Then, both the following must hold:*

$$(i) \sum_{v,u \in S, v \neq u} \Delta_{uv}^+ \leq 1$$

$$(ii) \sum_{v,u \in S, v \neq u} \Delta_{uv}^- \leq -\ell$$

**PROOF.** The first part follows directly from Theorem 5. For the second part, consider the following sequence of sets  $G_1, \dots, G_{\ell-1}$ :

$$\begin{aligned} G_0 &\stackrel{def}{=} \{p\} \\ G_1 &\stackrel{def}{=} \{p, w_1\} \\ G_2 &\stackrel{def}{=} \{p, w_1, w_2\} \\ &\vdots \\ G_{\ell-1} &\stackrel{def}{=} \{p, w_1, \dots, w_{\ell-1}\} \end{aligned}$$

For each  $G_i$ , the sum-sybilproofness condition shows that the total increment on edges crossing the  $G_i | \overline{G_i}$  cut is at most  $-1$ . Adding them all up, we have double-counted only edges with positive increments. With the double-counted edges, the total is  $-\ell$ ; thus, if each edge was counted only once, the sum would be at most  $-\ell$ .  $\square$

## 6. DISCUSSION AND FUTURE WORK

In this paper, we presented an abstract model of risky interactions in a community with pseudonymous users, and pairwise trust accounts. We analyzed protocols that allow chains of trust to be used when direct trust relationships are absent, and update trust values based on the declared outcome of each transaction. We presented a natural strategic property for such protocols, sum-sybilproofness, and showed that there is an unavoidable friction in using transitive trust methods in this setting: any protocol that is sum-sybilproof

must reduce the overall trust capital more aggressively on negative outcomes than it can increase the trust capital on positive outcomes, and thus, can lead to a reduction in long-run efficiency as potentially profitable interactions are foregone in the future. The bound on efficiency loss is tight: we showed that the hedged-transitive protocol and its variants are sum-sybilproof and they lose no more efficiency than required by the bound in Theorem 5.

Corollary 4 shows that the surplus of a transaction may be divided into three parts, while still maintaining the sum-sybilproof property. First, the principal might pocket some of the surplus and not place it at risk of future dissipation. That would be analogous to a gambler at a casino who begins the evening risking her initial chip buy-in but, after winning a few times, cashes out some chips in order to guarantee that some of those winnings will go home with her. Second, the principal might increase her trust balance for the intermediary. Third, the principal might increase the trust balance of the agent, so that in the future the agent might be trusted directly for a transaction. Increasing the trust balances exposes the principal to risk of future losses, should the intermediary or agent turn out to be sybils; however, it is needed to counteract the declining trust balances after bad outcomes, which can happen even with honest partners. One interesting direction for future research is to characterize the optimal distribution of transaction surplus to trust accounts, given assumptions about the prevalence and types of attackers.

Our analysis has focused on preventing damage from attackers when transitive trust protocols are used. Clearly, however, some reward may be necessary to induce voluntary participation from honest trust intermediaries. An increased trust balance from the principal might be of some utility to an honest intermediary, when there are potential future interactions between them that would be foregone if the principal did not have sufficient trust in the intermediary. In other cases, more direct reward of the intermediary, through money or gifts, may be appropriate.

Our definition of sum-sybilproofness embodies an ex-post notion of resistance to manipulation. This was based on our assumption that the outcome-generating process might itself be manipulated in arbitrary ways by an attacker controlling either the agent or the principal. If outcomes are generated by a process that is perfectly known to either  $a$  or  $p$ , a weaker *ex ante* notion of resistance may be adequate: honest agents bear some risk on each transaction, but the laws of large numbers apply to contain the aggregate long-term risk. In this case, in contrast to Theorem 5, there is no friction loss due to transitive transactions. However, in an online community in which value is generated through user interactions, it is more reasonable to assume that at least some aspects of the outcome-generating process are subject to influence by the interacting agents. It is therefore important to model situations in which there is partial knowledge of the distribution of outcomes, but there is residual uncertainty about this distribution. We conjecture that a modified version of Theorem 5 will hold in this case as well.

Finally, it would be good to examine broader classes of distributed trust accounting, instead of the pairwise trust accounts we study here: It may be possible to mitigate the threat of sybil attacks, and the resultant loss of efficiency, by tracking a reputation for groups of users in addition to individual users.

## Acknowledgement

This work was supported by the National Science Foundation under award IIS-0812042. We would like to thank Lada Adamic for valuable suggestions on applications of our result.

## 7. REFERENCES

- [1] A. Cheng and E. Friedman. Sybilproof reputation mechanisms. In *P2PECON '05: Proceeding of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems*, pages 128–132, 2005.
- [2] J. Douceur. The sybil attack. In *Peer-to-Peer Systems: First International Workshop, IPTPS'02*, volume 2429 of *Lecture Notes in Computer Science*, pages 251–260. Springer, 2002.
- [3] Epinions, 2009. <http://www.epinions.com>.
- [4] E. Friedman and P. Resnick. The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, 2001.
- [5] J. Hartline and T. Roughgarden. Optimal mechanism design and money burning. In *STOC '08: Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 75–84, New York, NY, USA, 2008. ACM.
- [6] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in P2P networks. In *Proceedings of WWW '03*, pages 640–651, 2003.
- [7] N. Kiyotaki and R. Wright. On money as a medium of exchange. *Journal of Political Economy*, 97(4):927–954, 1989.
- [8] N. Kocherlakota. Money is memory. Technical report, Federal Reserve Bank of Minnesota, 1996.
- [9] R. Landa, D. Griffin, R. Clegg, E. Mykoniati, and M. Rio. A sybilproof indirect reciprocity mechanism for peer-to-peer networks. In *Proceedings of IEEE Infocom '09*, 2009.
- [10] R. Levien. *Attack-Resistant Trust Metrics*. PhD thesis, University of California, Berkeley, 2004.
- [11] D. Reeves. Blog article, 2006. Available at <http://blog.yootles.com/2006/11/07/yootles-trivia-question/>.
- [12] D. M. Reeves, B. M. Soule, and T. Kasturi. Yootopia! *SIGecom Exchanges*, 6(2):1–26, 2007.
- [13] P. Resnick and R. Sami. The influence limiter: Provably manipulation-resistant recommender systems. In *Proceedings of the ACM Recommender Systems Conference (RecSys07)*, 2007.
- [14] P. Resnick and R. Sami. The informational cost of manipulation resistance in recommender systems. In *Proceedings of the ACM Recommender Systems Conference (RecSys08)*, 2008.
- [15] Ripplepay, 2009. <http://www.ripplepay.com>.
- [16] M. Yokoo, Y. Sakuri, and S. Matsubara. The effect of false-name bids in combinatorial auctions: New fraud in internet auctions. *Games and Economic Behavior*, 46:174–188, 2004.
- [17] P. Zimmermann. *Pretty good privacy: public key encryption for the masses*, pages 93–107. Springer-Verlag New York, 1995.