Computational Intelligence, Volume xx, Number 000, 20xx

USING IDENTITY PREMIUM FOR HONESTY ENFORCEMENT AND WHITEWASHING PREVENTION

LE-HUNG VU

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

JIE ZHANG

Nanyang Technological University (NTU), Singapore

KARL ABERER

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

One fundamental issue with existing reputation systems, particularly those implemented in open and decentralized environments, is whitewashing attacks by opportunistic participants. If identities are cheap, it is beneficial for a rational provider to simply defect when selling services to its clients, leave the system to avoid punishment and then rejoin with a new identity. Current work usually assumes the existence of an effective identity management scheme to avoid the problem, without proposing concrete solutions to directly prevent this unwanted behavior.

This paper presents and analyzes an incentive mechanism to effectively motivate honesty of rationally opportunistic providers in the above scenario, by eliminating incentives of providers to change their identities. The main idea is to give each provider an identity premium, with which the provider may sell services at higher prices depending on the duration of its presence in the system. Our price-based incentive mechanism, implemented with the use of a reputation-based provider selection protocol and a reverse auction scheme, is shown to significantly reduce the impact of malicious and strategic ratings, while still allowing explicit competition among the providers. It is proven that if the temporary cheating gain by a provider is bounded and small, and given a trust model with a reasonable low error bound in identifying malicious ratings, our approach can effectively eliminate irrationally malicious providers and enforce honest behaviour of rationally opportunistic ones, even when cheap identities are available. We suggest the identity premium function that helps such honesty sustained given a certain cost of identities and analyze incentives of participants in accepting the proposed premium. Related implementation issues in different application scenarios are also discussed.

Key words: Trust; Reputation; Multi-agents; Cheap Identities

1. INTRODUCTION

Reputation systems have been shown to be effective in enforcing honesty and facilitating trustworthy behavior in a variety of practical application scenarios. Prominent examples of these systems include business applications such as eBay, P2P content provisioning systems, service marketplaces, and social recommender systems, to name just a few. The effectiveness of a reputation mechanism in enforcing truthful behavior is due to its capability to detect and punish individuals with bad intentions (malicious and uncooperative), as such bad behavior results in low reputation as perceived by the community.

Pseudonyms such as nicknames are usually used to identify the participants in reputation-based systems. Generally, these pseudonyms are disassociated from real-life identities as a form of protecting the anonymity and privacy of the participating users. As a result, it is relatively easy for users to acquire and change their identities at a low cost. On the one hand, this disassociation is a must to facilitate interactions in online environments (Friedman and Resnick, 2001). On the other hand, it becomes possible for any (intelligent yet malicious) participants to *whitewash their bad reputation* and thereby effectively avoid punishment of the community after defection. This whitewashing issue is a fundamental problem in any reputation system: it is the main source of several attacks and vulnerabilities (Hoffman *et al.*, 2009). Marti and Garcia-Molina (2003) show via empirical simulation that the effectiveness of a reputation system compared to systems without any reputation mechanism varies largely depending on whether the system uses an easy-to-defect account registration (thus whitewashing

Address correspondence to LE-HUNG VU (lehung.vu@gmail.com) and JIE ZHANG (zhangj@ntu.edu.sg).

bad behavior is simple) to a hard-to-change account management scheme with permanent identifiers for users. The famous Sybil attack (Douceur, 2002) is also related to the problem of easy-to-change and cheap identities.

While the rich literature on trust management provides us valuable insights into the development of incentive mechanisms to enforce honesty in decentralized systems, there is little analysis on how to combat the problem of cheap pseudonyms and the whitewashing of bad behavior. Most work on trust and reputation models (implicitly) assume an underlying identity management infrastructure that handles the cheap pseudonym issue effectively and efficiently. Interested readers may want to refer to existing surveys to have a better view of the area (Hoffman *et al.*, 2009; Golbeck, 2006; Jøsang *et al.*, 2007; Despotovic and Aberer, 2006). In this work, we propose a dynamic pricing mechanism to create economic incentives for rationally opportunistic providers to stay in the system and use the same identity throughout their life-time, thereby effectively preventing their whitewashing behavior. Identities can still be easy to create, and no costly identity management approach is needed. Our solution is applicable in rational environments with opportunistic participants behaving strategically to maximize their expected life-time utilities. We also assume the existence of a few irrationally malicious providers whose goal is to attack the system at any cost.

Our incentive mechanism is designed as a protocol for a rational client to select the most eligible provider for a transaction. First, those providers offering services matching the client's requirements are checked if they ever defected in their most recent transactions. Specifically, the reliability of the most recent rating on each eligible provider is evaluated to determine whether the provider defected in the last transaction with a previous client. Well-experimented (reputation-based) computational trust models, e.g., those presented in (Xiong and Liu, 2004; Teacy *et al.*, 2005) can be used for this purpose. The evaluation of the last rating's reliability decides whether the provider is included for selection or blacklisted by the client. Interestingly, the consideration of only the most recent rating gives sufficient incentives for rational providers to cooperate in most of their transactions. It is proven that the protocol also helps to reduce the negative influence of the intentionally malicious participants if the dishonesty detector is accurate in identifying the unreliable and biased ratings. Second, those providers passing the evaluation are invited to participate in an anonymous reverse auction. The goal of the auction is to promote competition among providers and to discover the true price of the service. The auction winner, i.e., the one offering the lowest price, will be selected by the client for the next transaction with one important adjustment: the final price the client pays is determined based on the auction-winning price adjusted with the provider's identity premium. This identity premium is a function of the number of transactions associated with the provider's identifier and assures that well-established providers with good reputation will have a strong advantage against newcomers in terms of pricing their services even in competitive scenarios. Such an identity premium concept corresponds to what actually happen in practical business environments. As an example, a previous study (Resnick *et al.*, 2006) shows that buyers on eBay are willing to pay more than 8.1% the usual price to popular and reputable sellers. With the identity premium-based pricing model, any rational provider is given strong incentives to cooperate in all but its very last transaction, despite that cheap identities may be available.

Our proposed approach can be applied in various online reputation-based marketplaces with different degrees of centralization. Note that although most other systems use reputation information to determine a trustworthy provider in terms of providing service, we use reputation to evaluate both the trustworthiness of the provider and of the related services. As a result, this work provides the following contributions to the trust and reputation research community:

- We propose the use of identity premium as a way to provide strong honesty incentives for providers in open and decentralized environments with rational participants and cheap identities. We prove that an identity premium-based pricing model helps to ensure honesty from any rational provider in all but its very last transaction, and thereby the incentive of a provider to change its identity is effectively eliminated.
- We identify and analyze the relation amongst the effectiveness of the reputation-based computational trust model, the service pricing model, the cost of identities, and the provider's incentive of honesty. This contribution is significant since it provides fundamental understanding on to which extent a computational trust model can eliminate the intentionally malicious providers and enforce honesty of the rational ones, even when cheap pseudonyms are available.
- We analyze the incentives of providers and clients in accepting the proposed model via the analysis of a system using such an identity premium concept. We then identify those constraints and scenarios where such mechanisms are still beneficial and acceptable by clients and providers, depending on the application domain. We show that in any application, the problem of cheap identities can be solved if, in each transaction, the additional gain of a provider by cheating is limited. Even with the asymmetry of information between providers and clients,

using an identity premium-based pricing model only causes bounded reduction in revenue to any long-staying provider compared to an ideal system, thus it is still acceptable to them. Considering risks, our provider selection approach is preferable for clients compared to a system without any identity premium and where whitewashing attacks are likely to happen. Different possible approaches to implement such identity premium for providers are also discussed.

The rest of the paper is organized as follows. In the next section, we review the related work. We present the system model and introduce the basic concepts in Section 3. Section 4 presents in details our proposed approach and the corresponding analysis on the effectiveness of the solution. In Section 5, some issues on the implementation of the identity premium approach are discussed. We finally conclude the paper and propose future work in Section 6.

2. RELATED WORK

Existing solutions to the problem of cheap pseudonym and whitewashing behaviors are usually based on the principle of imposing a high entrance cost to the system. Among them, one possible implementation is to use strong, verifiable pseudonyms linked to real-world identities, thereby making it more difficult for users to switch their identities. For example, credit-card numbers or physical mailing addresses are required in some online auction sites, such as ebay.com or ricardo.ch, as a way to ensure that each person can open only one single account. Another alternative is to require any newcomer to pay a monetary fee when joining the system. These above solution can be effective in preventing whitewashing in small scale, centralized systems, but are very difficult to implement in a decentralized environment without any centralized authority. First, difficult-to-create identities and entrance cost may defer users' participation. It is also not trivial to implement the payment mechanism fairly in a decentralized setting, e.g., it is not easy to determine to whom newly joined participants should pay their entrance cost. Second, potential privacy leakage and anonymity breaches make the use of real-life identities in online systems very difficult, if not impossible.

Other work also proposes to assign a low initial reputation value to a new user to prevent similar re-entry problems, namely (Kerr and Cohen, 2010). In these works, the value of a high reputation is not studied explicitly in relation with the economic incentive of honesty of the rational participants. They also do not provide a detailed analysis and quantification to which extent such mechanisms limit the whitewashing intention of the opportunistic service providers. Furthermore, new users with low reputation values will have little chance to do business, and as a result it is difficult for them to build up their reputation. These solutions can be used as complementary to our identity premium approach. They allow users to see whether the low reputation of a provider is due to the fact that many of their interactions ended badly, or that they are new to the system. Thus, with the use of an identity premium, newly joined providers may sell their services at a potentially lower initial revenue and have the potential to build up their reputation.

Particularly, an early work addressing the issue of whitewashing behavior in reputation systems is presented in (Feldman *et al.*, 2004b, 2006), where adaptive stranger policies are used to treat newcomers depending on the behaviors of the past newcomers. Experiments showed that adaptive stranger policies work well with a reasonably small turnover rate of the users. This work is still preliminary since only simulation results on a simple P2P file sharing system are presented, without any further theoretical analysis and generalization for similar environments. Implementation of the above approach in a general open and decentralized system is unfortunately difficult, since the building of an effective adaptive stranger policy, as shown in (Feldman *et al.*, 2004b, 2006), requires a reasonably good estimate on the number of real newcomers and whitewashing malicious providers. Furthermore, in more critical business applications newcomers should be treated more carefully, and defecting has more serious effects and thus is less tolerant to the system's reputation. Feldman *et al.* (2004a) estimate the negative impact of whitewashers in a simulated P2P file system and conclude that by imposing penalties on all newcomers, whitewashers can be prevented. Our work can be seen as a possible decentralized implementation of the punishment for newcomers, and we have treated the subject more extensively by considering the impact of the computational trust model being used by the system, the identity cost of the participants, and the temporary cheating gain to the incentives for the rational providers to behave honestly.

Bhattacharjee and Goel (2005) analyze the relation between the reputation premium with the transaction costs the seller needs to pay to the centralized system to avoid ballot stuffing by a buyer in an eBay-like reputation system. No mechanism of using reputation premium for preventing whitewashing is presented.

Other potential approaches to deal with the easy-to-change identities problem are related to research efforts in entity resolution in the database research community (Shen *et al.*, 2005). The goal of these works is to identify whether many virtual pseudonyms refer to the same real person, thus fake identities can be detected and eliminated. Sybil-proof reputation mechanisms (Yu *et al.*, 2006; Resnick and Sami, 2009), which aim to detect identities of malicious users by investigating the structure of social links between the Sybils and honest users also help to prevent whitewashing behavior to a large extent. These approaches implicitly assume that the new participants expend a certain cost to build relationships with existing entities in the system. This is different from our approach, which considers a variety of scenarios where the identity cost may vary.

In the economics literature, there are several studies on the empirical phenomenon of premium. Most of these works present online field experiments showing the existence of reputation premium in a variety of marketplaces, e.g., in Bordeaux's wine market (Landon and Smith, 1998), and on auction sites such as eBay (Ba and Pavlou, 2002; Ghose *et al.*, 2009; Standifird, 2001). Among them the work most related to ours is (Shapiro, 1983), in which the authors quantify the premium that must be levied from sellers to enforce trustworthy behaviour and avoid the case of zero-cost identities. However, no detailed analysis of the case of noisy evaluation and possible malicious behaviors is given. Our analysis is also more general, as we study the case of identities with small non-zero cost more explicitly, and also suggest possible ways to implement such an identity premium in different application scenarios.

Our work can also be seen as a simple way of using reputation information to improve the trustworthiness of participants in online auctions. Thus this work is in some way related to existing studies on the best use of trust and reputation information to improve the efficiency and tackle trusting issues in electronic negotiations (König *et al.*, 2008).

3. SYSTEM MODEL

We consider a network of autonomous and intelligent agents participating in an online marketplace of services with the following operational constraints:

- Each agent has *a public identity* and can be a provider and/or client of a (possibly infinite) number of services. Agents may change their identities freely and inexpensively, i.e., an identity can be bought at a small cost $\xi \geqslant 0$.
- The system may be *decentralized*, yet there is a secure (decentralized) storage system that enables the reliable sharing of information among the agents, e.g., a storage system implemented on top of a Distributed Hash Table (Aberer *et al.*, 2003). Thus, we assume the existence of a shared public space implemented on top of the distributed storage layer for easy information sharing between any two agents.
- A provider sells its services, each at a prescribed quality level $q \in \mathcal{Q}$. If the provider claims to provide a service at a quality level q, with a price u_q, yet actually delivers it at a lower quality level $q' < q$, the provider is said to have defected. In that case the provider has an additional illegitimate gain $v_{qq'}$ due to its cost saving when delivering the lower quality service. If $q' = q, v_{qq} = 0$, i.e., the provider is said to be cooperative or honest. The temporary cheating gain is a function $v : \mathcal{Q} \times \mathcal{Q} \mapsto [0, v^*]$ bounded by some $v^* > 0$. In this work we only consider the most relevant case where the temporary cheating gain by a provider in a transaction is at most the price of the offered service, i.e., $v_{qq'} \leqslant v^* \leqslant u_q$. As the provider spends less to provide lower-quality services, $v_{qq'}$ is a monotonically decreasing function of q'. Therefore, a provider has incentive to offer services at a quality lower than it promises, i.e., to defect when providing the services. In this paper, we consider the case of two quality levels (good and bad), for which the index q is skipped and the service price is usually denoted as u. Similarly, we use v instead of $v_{qq'}$ for the enhanced readability of the paper and assume that $v = \gamma u$, where $0 \leqslant \gamma \leqslant 1$ is the temporary cheating gain ratio of a provider.
- After using the service, the client posts a binary rating on the quality of the transaction with the service provider. The ratings can be stored locally or globally, and any rating can be retrieved and verified to be authentic by any other agent in the system, e.g., by using existing cryptographical methods like digital signatures and digest hashes. For a given service, the transaction is considered as *good* (or +) iff the client perceives the service quality as good as it expects. Contrarily, the transaction is evaluated as *bad* (or -) by the client iff the service quality is lower than what promised by the provider. With our incentive mechanism, this binary rating system turns out to be able to enforce honesty of providers at the highest possible extent.
- We consider the case where most providers are *rationally opportunistic* in economics terms, i.e., they want to maximize their expected life-time utilities by behaving strategically in each transaction. A few providers are *intentionally (irrationally) malicious* and want to attack the system at any cost by defecting when delivering their services or by posting dishonest and biased ratings on their allies and competitors.

The above abstract model represents many reputation-based online marketplaces with different degrees of centralization. Such a model can represent, for instance, a centralized eBay-like auction site, a commercial trading system implemented on top of an online social network, or a decentralized market of computational or storage services (Papazoglou and Georgakopoulos, 2003; Buyya *et al.*, 2001). Consequently, our proposed solution can be used in all these applications.

4. SOLUTION FRAMEWORK

4.1. Fundamental Concepts

We suppose that each agent in the system uses a (reputation-based) computational trust model to evaluate the credibility of a rating on a provider with certain error rate. Thus, a client uses the trust management mechanism as a *dishonesty detector* to identify potentially malicious ratings, and to select a reliable provider for its future transactions (Def. 1). Note that in any reputation system, a provider may collude with some clients to create fake transactions in order to build up its reputation. Those ratings related to the fake transactions are also considered malicious by our definition and therefore the detection of them is also a part of the computation trust model mechanism.

Definition 1: A *dishonesty detector* $\mathcal{R}$ is a computational trust model that evaluates a rating as *reliable* or *unreliable*, using as input historical performance statistics of the provider, the rater and the other related agents. The detection procedure is *verifiable* by any agent.

The following statistical accuracy measures of a dishonesty detector, as presented in Def. 2, are of our interest.

Definition 2: We define the *accuracy of a dishonesty detector* $\mathcal{R}$ in estimating the reliability of a rating as the maximum *misclassification error* ε, where $0 < \varepsilon < 1$ and ε is common knowledge. ε is the upper bound for the actual misclassification rates α and β of the detector $\mathcal{R}$, corresponding to false positives and false negatives, i.e., $\alpha =$Pr(rating estimated as reliable by $\mathcal{R}$ | rating is actually unreliable), and $\beta =$ Pr(rating estimated as unreliable by $\mathcal{R}$ | rating is actually reliable).

The accuracy, or misclassification error bound, of a dishonesty detector also implies its resilience to possible malicious attacks from intelligent agents that manipulate ratings on their competitors and alliance. To be accurate, the computational trust model should consider the performance statistics of both the rater and the provider when estimating the reliability of a rating. Note that the actual value of α, β of a dishonesty detector may change (possibly improve) over time, depending on the learning capability of the detection algorithm. It is only necessary that the *upper-bound* ε of the misclassification errors of the dishonesty detection to be known and used as common knowledge in the system. This upper bound may be estimated through simulation-based or real-life experimentation as the prediction accuracy of the computational trust model in the presence of different types of misbehaviour.

In this paper, we assume the existence of a computational trust model as an effective dishonesty detector with a reasonable error bound $0 < \varepsilon < 0.5$, irrespective of the possible manipulation of ratings by the participants. Detailed implementation and analysis of such an evaluation mechanism is not the focus of this paper. However, we believe such an assumption is realistic due to the following reasons:

- An adversary may control a part but not all inputs to the dishonesty detection mechanism, e.g., a provider may collude with up to a certain percentage of the clients, so in general the adversary does not have strong influence on the detection error bound. This is also true since the detection algorithm can be open but its specific setting can be kept secret before the evaluation and thus gaming of the result by the involved agents can be avoided.
- Given a certain number of known (malicious) attack models and with the given limitation in the capability of the adversary, the designer can always come up with a sophisticated detection mechanism to detect these attacks and eliminate bogus ratings effectively.
- In real life, out of band monitoring and investigation mechanisms can also be used to learn the truthful outcome of a transaction with high accuracy, even though such an approach can be costly. Many existing reputation-based trust models in the literature can also be used to implement such a dishonesty detector with very high accuracy, as explained in detail in Section 5.2.

We propose the following procedure for a service client to select the best suitable provider among the eligible ones given their reputation, as summarized in Fig. 1. First, any provider s offering a service matching the client's requirements will be checked whether it defected in its most recent transaction. Def. 3 gives details of this evaluation. The providers passing this evaluation are then invited to participate anonymously in an online *reverse auction* to compete for the right to sell service to the client. Specifically, these providers will place their bid for the lowest service price they are willing to accept. The identities of the participants are hidden from each other. Such an anonymous auction promotes competition among providers and helps to discover the true price of the service (Schoenherr and Mabert, 2007). The identity sp of the auction winner, i.e., the one offering the lowest price, and its offering service price u, will be revealed after the auction. The final price the client pays for the selected provider sp is then determined based on the auction-winning price u adjusted with the identity premium of the winner. The identity premium of a provider is determined by the number of transactions it has finished in the system with its current identity (to be explained further in Section 4.3). The key idea is to give reputable providers a strong advantage against newcomers in terms of pricing their services, even in competitive scenarios. It is therefore beneficial for the provider to keep the same identity for future transactions during its stay in the system.

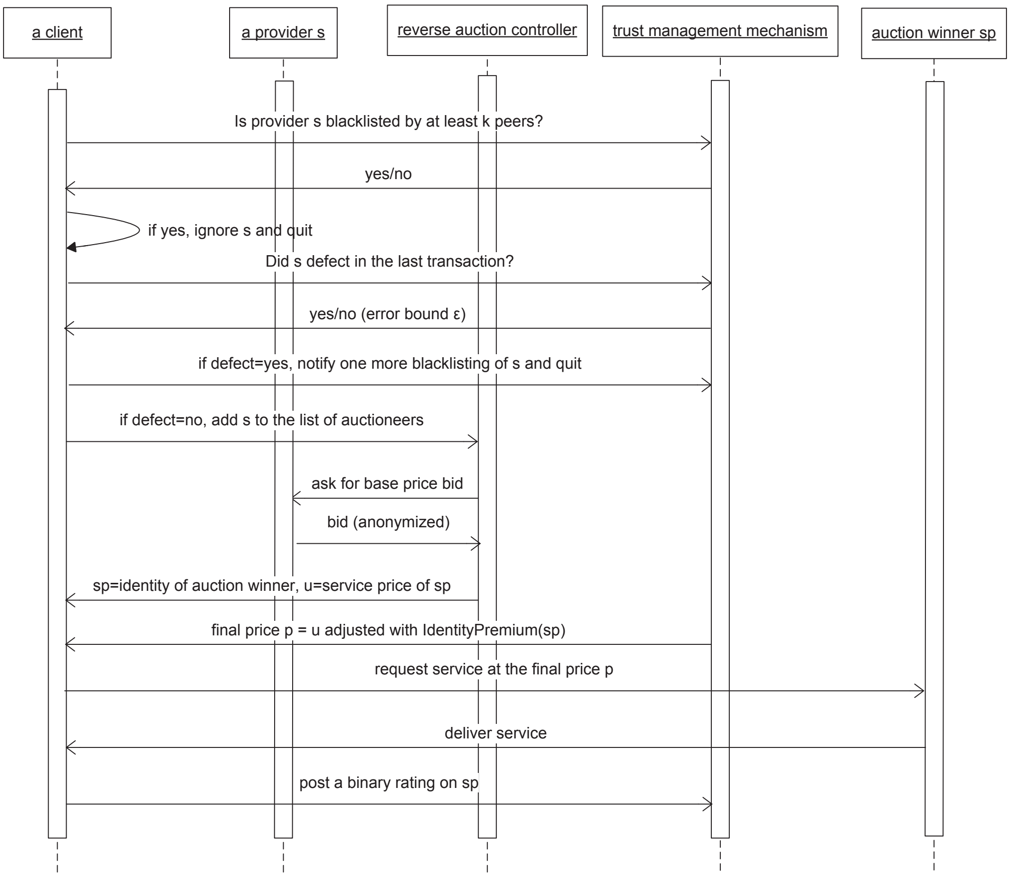

a client
a provider s
reverse auction controller
trust management mechanism
auction winner sp
Is provider s blacklisted by at least k peers?
yes/no
if yes, ignore s and quit
Did s defect in the last transaction?
yes/no (error bound ε)
if defect=yes, notify one more blacklisting of s and quit
if defect=no, add s to the list of auctioneers
ask for base price bid
bid (anonymized)
sp=identity of auction winner, u=service price of sp
final price p = u adjusted with IdentityPremium(sp)
request service at the final price p
deliver service
post a binary rating on sp

FIGURE 1. Different steps of selecting a reputable provider for the next transaction by a service client.

Definition 3: A client evaluates the eligibility of a provider with the following *provider selection protocol* $\mathcal{S}_k = \langle \mathcal{R}, k \rangle$:

(1) it retrieves the most recent binary rating $r \in \{+, -\}$ on the provider, considering the absence of a rating as the presence of a positive rating.

(2) the binary reliability $\widehat{t} \in \{\textit{reliable}, \textit{unreliable}\}$ of r is evaluated with the dishonesty detector $\mathcal{R}$.

(3) if $(\widehat{t} = \textit{reliable} \wedge r = -) \vee (\widehat{t} = \textit{unreliable} \wedge r = +)$, the client publishes this information (a detection of the most recent cheating by the provider) to the public storage space (a global blacklist).

(4) the provider is invited to the auctioning step if in the publicly shared space there are less than $k \geqslant 1$ published cheating detections on the provider regarding its most recent transaction. Otherwise the client blacklists this provider.

Essentially, Def. 3 specifies that for the selection of a provider, only the last interaction (from any client) with that provider is taken into consideration, and each client makes its own selection decision based on the outcome of the feedback on that last interaction. Note that the client may play the role of the auctioneer to choose the provider if required, e.g., in a fully decentralized system.

TABLE 1. Commonly used notations

Notation	Definition
u_*	minimal price of offered services, $u_* > 0$
u^*	maximal price of offered services, $u^* \geqslant u_*$
u	the service price proposed by the reverse auction winner, $u_* \leqslant u \leqslant u^*$
v	the cheating gain of provider if it defects, $0 \leqslant v \leqslant u$
u_i [or v_i]	similar to u [or v] but considered in the context of the i-th transaction
ξ	the cost to create a new identity
ξ_0	the minimal identity cost $\gamma\lambda(1-\phi)$ to enforce honesty of providers
γ	the ratio v_i/u_i [or v/u]
$\mathcal{R}$	a dishonesty detector to estimate reliability of a rating, as explained in Definition 1
α	Pr(rating estimated as reliable | rating is actually unreliable)
β	Pr(rating estimated as unreliable | rating is actually reliable)
ε	upper-bound of α and β, $0 < \varepsilon < 0.5$
k	# of posted cheating detection on a provider, before it is globally blacklisted by clients
λ	the relative cheating gain ratio $\gamma/((1-\varepsilon)^k - \varepsilon^k)$
$\mathcal{S}_k = \langle \mathcal{R}, k \rangle$	a provider selection protocol specified in Def. 3
Δ	# of the remaining services of a provider
ϕ	a parameter determining the initial price of a service, $0 < \phi < 1$, potentially with index i referring to a certain service in the i-th transaction of a provider

Fig. 1 shows the step-by-step illustration of the above provider selection protocol. Such a protocol is tough for bad providers, including malicious and rationally opportunistic ones. It assures that a globally blacklisted provider has to quit the system and joins in with a new identity if ever wanting to sell its services again. The reverse auction controller and the trust management mechanism (i.e., the dishonesty detector) can be implemented as a centralized entity or in a decentralized way at each client depending on the degree of centralization of the system.

The evaluation of rating reliability by a dishonesty detector in step (2) of Def. 3 helps to reduce the influences of strategic rating manipulation by rational or malicious agents. The goal here is to eliminate as many malicious providers as possible when they start cheating and incentivize rationally opportunistic providers to cooperate. Actually, the use of the above selection protocol with a global computational trust model mimics the behavior of a centralized reputation system in practice. The parameter $k \geqslant 1$ represents the cautiousness of a client in trusting cheating detections published by others. For easy reference, Table 1 summarizes the most frequently used notations in this paper.

4.2. Scope of Our Analysis

To reduce the complexity of the analysis and the presentation clarity without reducing the applicability of the approach, we do not consider the following issues in our analysis. These issues are either orthogonal to the current problem or can be readily resolved with existing known solutions.

First, we do not directly consider the incentives of the clients to leave a rating after a transaction. Intuitively, the clients have indirect incentives to leave reliable ratings after transactions since this helps to eliminate bad providers. Also, it is possible to integrate existing incentive mechanisms, e.g., via side-payment (Miller *et al.*, 2005; Zhang *et al.*, 2012), to motivate the clients to leave honest feedback after their transactions. Furthermore, the absence of a rating after a transaction is considered as the presence of a positive rating, thus appropriate decisions can still be made even in case few ratings are available.

Similarly, providing incentives to share the result of the learning step (the evaluation of the rating reliability) is an orthogonal issue to the current analysis. This issue is in fact less relevant: in case the others do not share their detection results, a client can still do the detection by itself. As the learning at step (2) of Def. 3 is verifiable, (maliciously) writing of wrong learning results is detectable and not an issue. With $k = 1$ we even do not need the condition that $\mathcal{R}$ is verifiable as in Def. 1.

One possible problem with the protocol in Def. 3 is the potential badmouthing attack, when many agents collude to badmouth a certain provider. To reduce the effect of this attack and potential observation noise, a robust detection algorithm $\mathcal{R}$ should be used to consider the trustworthiness of both the rater and the provider when estimating whether a rating is reliable. As we will analyze, the accidental blacklisting of a good provider is of no harm to a buyer. Even if it may cause certain harm to a provider, such mistakes can be reduced by both increasing k and lowering the error bound of the second step by using a more expensive and sophisticated trust model. Designing such an accurate computational trust model is, however, not the focus of this paper.

4.3. Honesty Enforcement with Cheap Identities

We proved in the early work (Vu and Aberer, 2011) that if identities are costly, the selection protocol in Def. 3 assures that any rational provider is motivated to cooperate in all but a small number Δ_v of its last transactions. This Δ_v is dependent on the misclassification error bound ε of the trust mechanism being used in the system to detect malicious ratings and the bound v^* of the temporary cheating gain of a provider after a transaction.

The main goal of this paper is to develop an approach to enforce honesty even if identities are cheap. Our proposed solution is to use an *identity premium* (Def. 4) to determine the price of a service. This premium allows a provider to sell its services at higher prices depending on the number of transactions it has completed in the system with the current identity. Under this pricing scheme, an initially cheap pseudonym would have an increasingly significant value over time, thereby effectively eliminating the incentive of switching identities and whitewashing bad behavior of any opportunistic provider. As a result, almost full honesty is the best response strategy of any rational provider when participating in a transaction.

Definition 4: A provider agent that has finished $L > 0$ transactions using a given identity has a monotonically increasing *identity premium* $f(L)$ associated with that identity, where $f(0) = 0$. That is, consider a service with a base price $u_* \leqslant u \leqslant u^*$ (u is the winning price of the reverse auction in Fig. 1). A client pays the provider with the final price $P(\phi, f)$ where:

$$P(\phi, f) = u(1 - \phi) + f(L) \tag{1}$$

The price at no identity premium, i.e., with $f(0) = 0$, is determined by a parameter $0 < \phi < 1$, possibly depending on the base price u.

The parameter ϕ is set individually by each client in the system and determines the price that a newly joined provider can charge, so that this new provider has to sell its services with lower prices at the beginning. Introducing the parameter ϕ also offers the flexibility of setting higher prices by the identity premium for a provider staying in the system and behaving honestly for many transactions. The lower prices at the beginning thus will be compensated by the higher prices later on. Conditions of this parameter will be investigated in Theorem 1. The parameter ϕ also provides the flexibility for system designers to minimize system inefficiency such that both providers and clients are willing to accept our identity premium-based approach (see further analysis in Sections 4.4 and 4.5).

According to Def. 4, a client agrees to pay a provider having completed many transactions with a higher price. This price premium is built on the proven track record of the provider and thus closely related to the reputation score of the provider. The reason we chose the number of transactions L associated with an identity to determine the premium is due to its verifiability. On the other hand it is non-trivial to estimate the reliability of a reputation value: reputation may be estimated in a personalized manner and subject to various strategic manipulation by intelligent agents. An identity premium can be implemented in different ways, as discussed later in Section 5.

In this section we will study the properties of the identity premium function $f(L)$ to achieve the highest possible honesty from the opportunistic service providers in relation with the identity cost ξ and under the presence of strategic or malicious manipulation of ratings by competing providers.

Apparently, under the pricing scheme $P(\phi, f)$, a newly joined provider must sell a service at a price $u(1 - \phi)$, lower than the base price u of the service. A provider staying in the system for many transactions may sell services at higher prices in later transactions. Therefore staying in the system with the same identity helps the provider to compensate its loss during earlier transactions where it has zero or small identity premium. Therefore, even if establishing new identities is relatively cheap, it is expected that every rational provider finds it optimal to keep the same identity for the whole life-time rather than cheating, leaving, and joining with a new identity. In other words, whitewashing is not optimal for any rational provider.

Recall that the temporary cheating gain when a rational provider sells a service of any price u as γu, where $0 < \gamma \leqslant 1$ (Section 3). For presentation clarification, denote $\lambda = \frac{\gamma}{(1-\varepsilon)^k - \varepsilon^k}$. For simplicity, we call λ the *the relative cheating gain ratio* of a provider in a system with a given dishonesty detection capability ε. Given a bound ε for α, β of the dishonesty detector being used by peers in the system, Theorem 1 shows the relation between the error bound ε, the identity cost ξ, the identity premium function $f(L)$, and their effectiveness in enforcing honesty of a provider during its life-time in the case with possible cheap pseudonyms. The proof is available in the appendix of the paper.

Theorem 1: Assume that every client uses the protocol $\mathcal{S}_k = \langle \mathcal{R}, k \rangle$ where the dishonesty detector $\mathcal{R}$ has the misclassification errors α, β upper-bounded by $\varepsilon < 0.5$, to select a reputable provider for the next transaction.

Consider any rational provider with N services to sell. Let $u_* \leqslant u_i \leqslant u^*, i = 1, ..., N$ be the base prices of the services in the i-th transaction, as bid by the winning provider in the reverse auction. Suppose that the pricing model $P(\phi, f)$ is used by the client, and it follows that:

(i) If the identity cost ξ is small, the following identity premium ensures that honest behaviour is always the best response strategy of the provider in *any* transaction $i = 1, ..., N-1$, for any $0 < \phi_i < 1$:

$$f(L) \quad = \quad \sum_{i=1}^{L} \lambda^{L-i}(\lambda u_i(1-\phi_i) - \xi/\gamma) \text{ for } L > 0 \tag{2}$$

(ii) For $\lambda \neq 1$, let us consider the case of no competition among providers, where the base price u_i's can be assumed as a unit cost: $u_i = 1$ and $\phi_i = \phi, i = 1, ..., N$. If the identity cost $\xi < \xi_0 = \gamma\lambda(1-\phi)$, the following identity premium function is sufficient to enforce honesty for a provider in selling every but the last service:

$$f(L) \quad = \quad ((1-\phi)\lambda - \xi/\gamma)\frac{1-\lambda^L}{1-\lambda} \text{ for } L > 0 \tag{3}$$

With $u_i = 1, \phi_i = \phi, i = 1, ..., N$, and for $\lambda = 1$, the identity premium is:

$$f(L) \quad = \quad L(1 - \phi - \xi/\gamma) \text{ for } L > 0 \tag{4}$$

(iii) Let N_h be the number of transactions that a fully cooperative (honest) provider can participate till it is mistakenly blacklisted, and let N_c be the number of bad transactions an intentionally malicious provider can benefit from defecting until being eliminated from the system respectively. We have $E[N_h] > 1/\varepsilon^k$ and $E[N_c] < 1/(1-\varepsilon)^k$.

The results (i,ii,iii) hold even in presence of strategic manipulation of ratings by agents.

The $f(L)$ defined in Theorem 1 determines the additional amount the client must pay to motivate the provider. This additional cost of the client to motivate honesty of rational providers in the system is an inevitable cost in any open system with cheap identities, as proven in (Friedman and Resnick, 2001).

Apparently, a smaller identity premium $f(L)$ results in lower prices and thus is more encouraging to the clients. A provider, on the other hand, must sell its first service at a low price to gain the identity premium, and gets compensated by selling other services in the future transactions at higher prices. If providers have many services to sell, eventually they would be able to sell their services at very high price thanks to their accumulated identity premiums. As from (3), small values of $\lambda = \frac{\gamma}{(1-\varepsilon)^k - \varepsilon^k} < 1$ are preferable, since the prices with premium are not getting extremely high, thus still acceptable by the clients. We will investigate the different options of rational agents (clients/providers) and identify those conditions under which it is still beneficial for them to accept such an identity premium-based pricing approach in Section 4.4 and Section 4.5.

The service price ad infinitum, determined by the identity premium, is strongly dependent on *the relative cheating gain ratio* λ. This λ is decided by the characteristics of the marketplace, namely the additional cheating gain in a transaction $0 < \gamma = \frac{gain}{price} \leqslant 1$, the error bound ε of the trust mechanism being used to detect unreliable ratings, and the system threshold k to blacklist ill-behaved providers.

For $\lambda \geqslant 1$, the price of services ad infinitum cannot be bounded and depends on the number of services the provider wants to sell during its whole life-time. For $\lambda < 1$, it is possible to use an identity premium so that the price ad infinitum $L \to \infty$ is bounded as follows. If $u_i = 1, i = 1, ..., N$, as in item (ii) of Theorem 1, it is clear that $\lim_{L\to\infty} f(L) = \frac{\lambda(1-\phi)-\xi/\gamma}{1-\lambda}$. So, given ε, k, γ such that $\lambda < 1$, we can easily find the initial price $1 - \phi$ such that at infinity the identity premium-based service price reaches the standard price $u_i = 1$. In fact:

$$1 - \phi + \frac{\lambda(1-\phi) - \xi/\gamma}{1-\lambda} \quad = \quad 1 \Leftrightarrow 1 - \phi = 1 - \lambda + \xi/\gamma \tag{5}$$

We can also obtain a similar result for the case where u_i's are different by using the fact that $u_i \leqslant u^*, i = 1, ..., N$, letting $\phi_i = \phi$, and considering the following alternative for the identity premium in (2):

$$f(L) \quad = \quad (\lambda u^*(1-\phi) - \xi/\gamma)\sum_{i=1}^{L}\lambda^{L-i} = (\lambda u^*(1-\phi) - \xi/\gamma)\frac{1-\lambda^L}{1-\lambda} \tag{6}$$

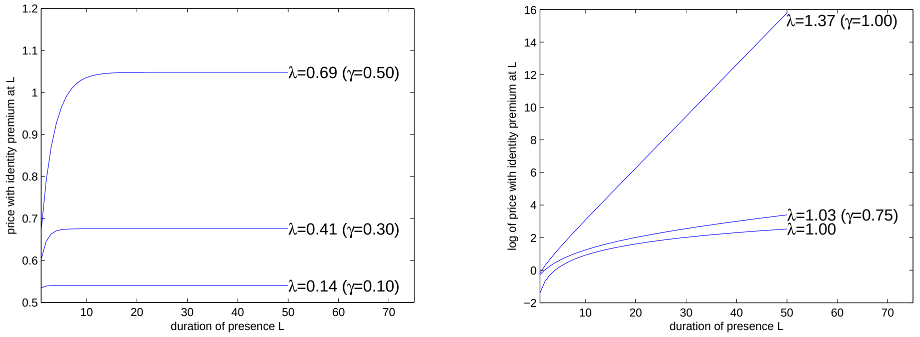
Under the identity premium described in (6), if $\lambda < 1$, the price determined by the identity premium of a provider can be understood as the price a client is willing to pay given its estimate of the defection probability of the provider. This can be shown by rewriting the price $P(\phi, f)$ as:

$$P(\phi, f) \quad = \quad u(1-\phi) + f(L) = u[1 - (\phi - \frac{f(L)}{u})]$$

As $\lim_{L\to\infty} \frac{f(L)}{u} = \frac{u^*\lambda(1-\phi)-\xi/\gamma}{u(1-\lambda)}$, the term $\phi - \frac{f(L)}{u}$ can be seen as a probability iff:

$$0 < \phi - \frac{f(L)}{u} \leqslant 1 \quad \Leftrightarrow \quad \lim_{L\to\infty}\frac{f(L)}{u} \leqslant \phi \Leftrightarrow \phi \geqslant \frac{1 - \frac{\xi}{\gamma\lambda u^*}}{1 + \frac{u(1-\lambda)}{u^*\lambda}} = \phi_{min}$$

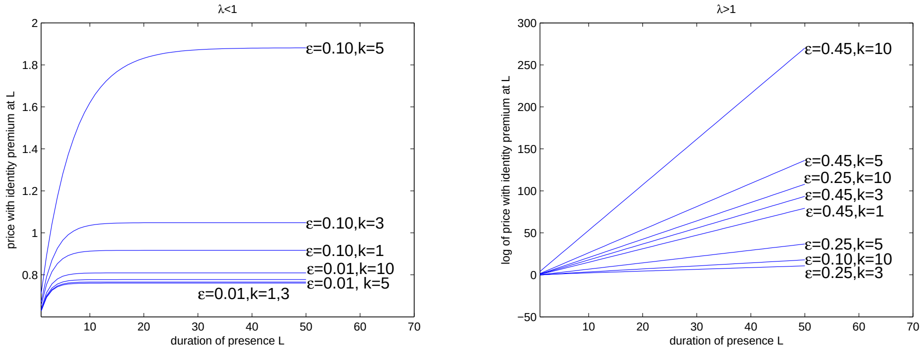
Seeing the price $u[1-(\phi-\frac{f(L)}{u})]$ of the service in a transaction as an average price based on the probability of defection by the provider, this probability could be estimated as $\phi - f(L)/u$. Thus, the longer a provider stays in the system, the smaller this probability of defection becomes, and the closer the price reaches the standard price u of the service.



(a) (b)

FIGURE 2. The identity premium-based price with different relative cheating gain ratios (a) $\lambda < 1$ and (b) $\lambda \geqslant 1$.

Fig. 2(a) and Fig. 2(b) show the identity premium-based prices for different values of the relative cheating gain ratio λ by a provider. The result is shown with the standard price $u = 1$ for a service, the initial price $1 - \phi = 0.5$ (half of the standard one), and the dishonesty detection mechanism setting $\varepsilon = 0.1, k = 3$. The identity cost is set at $\xi = \xi_0/2$. As shown in Fig. 2(a), the identity-premium price is bounded and reasonably small for $\lambda < 1$. For $\lambda \geqslant 1$ in Fig. 2(b) the price reaches high values very rapidly, which means an identity premium-based pricing model is not an option for systems with very high temporary cheating gains γ for rational providers. Further analysis (not given here) also shows that the starting price $1 - \phi$ and the identity cost ξ do not have significant influence to the price if $\lambda > 1$. With small $\lambda < 1$, where our premium approach is acceptable to clients, a higher identity cost ξ does help to adjust the price with identity premium considerably.



(a) (b)

FIGURE 3. The identity premium-based prices with different ε, k such that $\lambda < 1$ (a) and $\lambda > 1$ (b) compared to the standard price $u = 1$. The case $\lambda = 1$ is already shown in Fig. 2(b).

The effect of the error bound ε of the dishonesty detector to the identity premium-based price is given in Fig. 3(a) and Fig. 3(b), for an example case with $u = 1$, $\phi = 0.5$, $\gamma = 0.5$ and $\xi = \xi_0/2$. It is observed that the smaller ε (the more effective the dishonesty detection), the smaller λ becomes and the lower the price with identity premium. Given a specific γ, higher values of ε or k may result in $\lambda \geqslant 1$ and thus are not preferable. Thus for each application-dependent setting γ, there is an upper-bound of the misclassification error ε of the computational trust model used to implement the dishonesty detector, with which the identity premium is bounded and still acceptable to the clients.

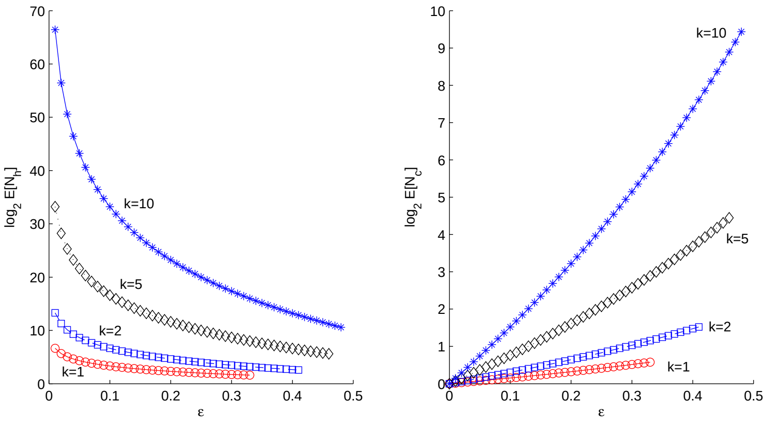


FIGURE 4. The relation between the misclassification error bound ε of a computational trust model and the *worst case lower-bound* of $E[N_h]$ (left side). On the right hand side is the relation between ε and the *worst case upper-bound* of the $E[N_c]$.

A thorough analysis of the effect of the intentionally malicious providers to the honesty among participants is out of the scope of this paper. Nevertheless, this effect is quantified by the claim (iii) of Theorem 1: in a system using a dishonesty detector with a small error bound ε, an intentionally malicious provider can only cheat for a limited number of transactions with one identity. On the other hand, fully cooperative providers are less likely to be mistakenly blacklisted and thus can use the same identity for many transactions. Fig. 4 shows the possibility of correctly eliminating the malicious providers and wrongly blacklisting the honest ones versus different ε and k for $\gamma = 1$. It is observed that the higher the accuracy of the computational trust model being used (lower ε), the lower the probability an honest provider is accidentally blacklisted (higher $E[N_h]$) and the higher the probability malicious providers are eliminated from the system (lower $E[N_c]$). Higher thresholds k reduce the possibilities of wrongly blacklisting honest providers, yet also increase the survival chance of intentionally malicious providers. Hence, given a known ε, it is recommended to choose appropriate k values given the prior information on environment vulnerability. In environments with more malicious behaviors, it is better for rational clients to choose smaller k values to eliminate bad providers quickly, at the cost of ignoring good providers. In less vulnerable environments with most good providers, higher k is recommended. Note that the given trends are for the worst case scenario in an extremely vulnerable environment, where an honest provider is repeatedly badmouthed by other users, and a malicious provider has enough resources for disguising its cheating activities by posting many positive ratings to the system consecutively. As we consider only a few of these irrationally malicious providers, the honesty level in the system should not be significantly affected. Of course, the system is still vulnerable to the attacks by several intentionally malicious providers that join the system with new identities and continuously defect to destroy the system reputation. These attacks cannot be prevented, yet one may consider that these intentionally malicious providers may have only limited resources, and thus the impact of these bad providers may be quantified in relation with this cost limitation.

4.4. System Inefficiency and Incentives of Providers to Accept an Identity Premium-based Pricing Mechanism

A key issue in the application of our identity premium-based approach in practice is the acceptance of the participants. Towards this answer, we will first analyze from the perspective of a provider its rational incentive to accept an identity premium approach.

Under the pricing scheme $P(\phi, f)$, a new provider must sell a service at an initial low price, and then gain premiums over time to compensate for its previous losses. It is important to understand whether a provider may gain or lose significantly from such a pricing scheme. Compared to an ideal case where a provider may sell every of its services at a competing price, the additional benefit $g(N)$ that the pricing scheme $P(\phi, f)$ gives a provider with N services to sell is defined as:

$$g(N) = \sum_{i=1}^{N}[u_i(1-\phi_i) + f(i-1)] - \sum_{i=1}^{N} u_i = \sum_{i=1}^{N}(f(i-1) - u_i\phi_i) \tag{7}$$

where $u_* \leqslant u_i \leqslant u^*$ is the base price of the service sold in i-th transaction (the winning price of the auction), ϕ_i is the parameter deciding the initial price at zero identity premium, and $f(i)$ is the identity premium of the i-th transaction.

Depending on the nature of the problem and the sequence of services sold by a provider, the total gain $g(N)$ of the provider from its identity premium can be positive or negative. If $g(N) > 0$, in the current system clients pay higher prices for services compared to a system without an identity premium-based pricing mechanism. As a result, higher values $g(N)$ may deter clients from participating in the system, as services are generally more expensive. However, as in the subsequent analysis, the case $g(N) > 0$ might still be acceptable to clients as their risk is reduced. On the other hand if $g(N) < 0$, providers must accept selling services at lower prices compared to normal systems where no identity premium are used. Negative values of $g(N)$ may deter participation of providers since they generally get less revenue.

Collectively, the average gain $g(N)$ reflects the inefficiency of the system using an identity premium-based pricing scheme $P(\phi, f)$. This inefficiency strongly affects the incentives of participation of clients and providers. $g(N)$ is the cost inherent to the cheap identity issue, which we will try to minimize by finding the optimal system design parameters, e.g., ϕ_i, given other fixed, domain-dependent variables, including the base service prices u_i's and the cheating gain ratio γ.

Theorem 1 puts a requirement on the shape of the identity premium function $f(.)$, yet it places no special requirements on the parameter ϕ that decides the initial price of a service. Therefore, we would try to find an optimal value of ϕ that determines the initial service price so that the inefficiency of the system is as small as possible.

Theorem 2: Consider a simple case where providers sell services of the same base price of 1. We have:

- With $\lambda \neq 1$, for any provider with N services to sell and N is known, the identity premium-based pricing mechanism has no inefficiency to the system w.r.t this provider, i.e., $g(N) = 0$, if the initial price is set at $1 - \phi$ where:

$$\phi \quad = \quad \frac{(\lambda^N - N\lambda + N - 1)(\lambda - \xi/\gamma)}{\lambda^{N+1} - (N+1)\lambda + N} \tag{8}$$

- If providers have a large but unknown number of services to sell and $\lambda < 1$, due to the asymmetry of information between providers and clients, the identity premium-based pricing mechanism has a small bounded inefficiency of $-\frac{\lambda-\xi/\gamma}{1-\lambda}$ to each provider if the initial price for a service approximates $1 - \lambda + \xi/\gamma$.

Proof. For $\lambda \neq 1, u_i = 1, i = 1, ..., N$, the identity premium function is given by:

$$f(L) \quad = \quad ((1-\phi)\lambda - \xi/\gamma)\frac{1-\lambda^L}{1-\lambda} \tag{9}$$

The inefficiency of the pricing mechanism to the provider is:

$$g(N) \quad = \quad \sum_{i=1}^{N}(f(i-1) - \phi) = \sum_{i=1}^{N}(((1-\phi)\lambda - \xi/\gamma)\frac{1-\lambda^{i-1}}{1-\lambda} - \phi) \tag{10}$$

$$= \quad \frac{(\lambda^N - N\lambda + N - 1)(\lambda - \xi/\gamma) - \phi(\lambda^{N+1} - (N+1)\lambda + N)}{(1-\lambda)^2} \tag{11}$$

$$g(N) = 0 \quad \Leftrightarrow \quad \phi = \frac{(\lambda^N - N\lambda + N - 1)(\lambda - \xi/\gamma)}{\lambda^{N+1} - (N+1)\lambda + N} = \frac{A}{B} \tag{12}$$

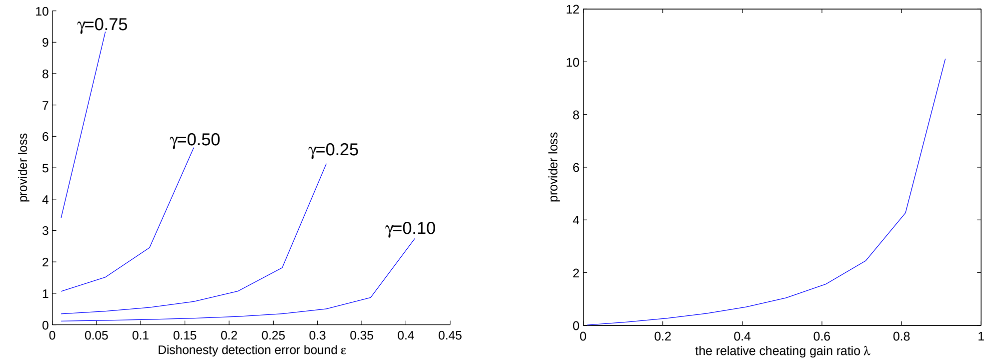
It can be verified for any $\lambda > 0$, $A - B < 0$. Therefore, with any $\lambda > 0$, from (12) we always have $0 < \phi < 1$.

For N very large and unknown to clients and $\lambda < 1$, any client can estimate the optimal setting $\phi \to \frac{-\lambda^2+\lambda+\lambda\xi/\gamma-\xi/\gamma}{1-\lambda} = \lambda - \xi/\gamma$, thus a client accepts to buy service from a provider at an initial price $1 - \lambda + \xi/\gamma$. The price ad infinitum in this case is $(1 - \xi/\gamma)(1 + \xi/\gamma/(1-\lambda))$. With this specific setting, the inefficiency of each provider is $\lim_{N\to\infty} g(N) = -\lim_{N\to\infty} \frac{(\lambda-\xi/\gamma)(\lambda^N-\lambda^{N+1}+\lambda-1)}{(1-\lambda)^2} = -\frac{\lambda-\xi/\gamma}{1-\lambda}$. □

In applications where $\lambda \geqslant 1$, it is apparent that providers have no objection to an identity premium approach: providers can sell services at very high prices compared to the standard values in latter transactions, and therefore gain much higher benefits. Therefore, we only need to consider the incentives of rational providers to accept an identity premium-based pricing approach in the nontrivial case $\lambda < 1$. Our conjecture is that the providers will still be better off accepting such an approach, for the following reason: in a system not using identity premium, where all participants are fully rational and several Nash equilibria may co-exist (Huang *et al.*, 2013), the system may converge to the worst Nash equilibrium, since providers would always cheat and therefore no client buys any service from any provider. As a result, a rational provider would gain nothing from selling its services. On the other hand, given a system using an identity premium pricing model with $\lambda < 1$, even though initially providers have to sell service under standard price, it is still beneficial for providers. However, more detailed analysis and experiments have to be done to confirm or reject this conjecture in future work.

According to Theorem 2, Fig. 5(a) shows the loss of a rational provider that sells infinitely many services (with the standard price of each service $u = 1$) versus the error bound ε of the dishonesty detection mechanism (the effectiveness of the trust measure being used). This loss is merely due to the information asymmetry between providers and clients: the number of services of a provider is unknown to client. We consider a representative case with a zero identity cost $\xi = 0$ and $k = 3$. Firstly, we observe that for an application with a given γ, there is an upper bound of ε so that $\lambda < 1$ and thus the price ad infinitum is upper-bounded and small. For smaller temporary cheating gains γ, this upper bound is higher, and thus the system is more error-tolerant in identifying malicious ratings. Secondly, it is apparent that the higher the effectiveness of the trust mechanism used (lower ε) and the lower the temporary cheating gain γ, the lower the loss of the provider (compared to the service price of 1). For example, at $\varepsilon = 0.25, \gamma = 0.25$, the provider's loss is approximately 1.5, i.e., if the provider participates in many transactions, its loss due to our pricing mechanism, termed the social cost of cheap pseudonyms as in (Friedman and Resnick, 2001), is 1.5 of the price of a single service. This loss is in fact very small and well acceptable to the provider, compared to the potential loss of the provider in a system with no identity premium, where rational clients distrust and do not buy any of its services. Fig. 5(b) shows a similar trend in the relation between the provider's loss and the relative cheating gain ratio $\lambda < 1$. In general, a rational provider will find good incentives to accept the price model with an identity premium concept, since its loss due to selling service at smaller prices upon joining the system will be almost compensated in its later transactions.

It is worth noting that there may be many factors or unavoidable circumstances (for instance, where the



(a)

(b)

FIGURE 5. The loss of a provider given an identity premium-based pricing model vs. the dishonesty detection error bound ε ($k = 3$) at different values of (a) γ; and (b) λ.

transportation company made a mistake) causing an honest negative rating of a provider. This may cause the provider to be blacklisted and force the provider to create a new identity and go through the initial cost of establishing this identity. In consequence, this may change the incentives for a provider to accept our identity premium mechanism. We have shown in our previous work (Vu and Aberer, 2011) that honest negative ratings due to unavoidable circumstances have only minor impact on providers.

4.5. Incentives of Rational Clients to Accept Identity Premium

As shown during the previous analysis, for those applications where $\lambda < 1$, the identity premium-based service price ad infinitum is bounded. Furthermore, the initial price can be set such that at infinity the price with identity premium is close to the standard base price u. This means rational agents, as potential clients, would be willing to accept such a pricing approach. For scenarios with $\lambda > 1$ and where providers may sell a limited number of services, as claimed in Theorem 2 (i), it may still be possible to set the initial price such that there is no advantages to providers. That means, in average, agents as clients for many transactions still find incentives in joining the system.

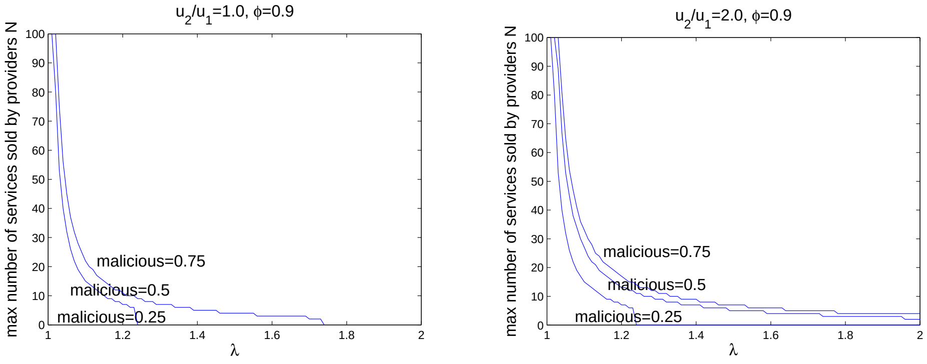
Another reason why agents, as clients, may accept the identity premium-based approach comes from the observation that in general, clients are risk-sensitive and favor systems with lower transactional risks. Let us compare the risk of clients when participating in two systems: the first one with an identity premium-based pricing scheme, and the second without such a mechanism. By convention, the risk of a client in buying a service with a price p and with a probability c that the provider cheats in the transaction is defined as pc.

We consider the following two cases, the first case is when the providers sell infinitely many services. According to Theorem 1, the probability that a provider (in a system using an identity premium) cheats in any transaction is zero. Thus the risk of a client when participating in a system with identity premium-based pricing is also zero, irrespective of the price of the service. This is true for every client in any transaction and for any $\lambda > 0$.

In the second case, suppose that each provider sells a limited number of services, and the distribution of number of services N of providers is a known cumulative distribution function $F(N)$. From Theorem 1, in a system where the service price is determined based on a provider's identity premium, a rational provider would only defect in selling the very last service. Consider the transaction on a service with the base service price u_1 offered by an opportunistic provider that has finished L transactions. The probability the provider defects would be the probability that this provider has no more services to offer after the current transaction, which is $c = F(L) - F(L-1)$. The risk of the client in transacting with this provider is thus:

$$risk_1(u_1, L) \quad = \quad u_1(1 - \phi + \frac{f(L)}{u_1})(F(L) - F(L-1))$$

In a system without identity premium and with cheap identities, let $m, 0 \leqslant m \leqslant 1$ be the general whitewashing intention of an opportunistic provider. The parameter m may represent the malicious turn over rate of these rational participants, or the probability that they defect and switch identities when an effective



(a) (b)

FIGURE 6. The maximal number of services sold by providers N for different λ with which risk-sensitive clients find an identity premium-based price model a better option. The relation is plotted with varied probabilities of detection m by providers and with (a) $u_2/u_1 = 1, \phi = 0.9$ and (b) $u_2/u_1 = 2, \phi = 0.9$

identity management scheme has not been implemented. Let the base service price a provider offers to its clients be u_2, then it is reasonable to assume that $u_2 > u_1$, since in a system where providers having no identity premium, they have no reason to offer a price lower than what they can afford. All other factors (such as the rationality of the provider and its malicious intention) being equal, the risk of the client in doing the current transaction in a system without any identity premium is:

$$risk_2(u_2, L) = u_2 m$$

To analyze the potential benefits of the clients in accepting an identity premium-based pricing model, we compare $risk_1(u_1, L)$ and $risk_2(u_2, L)$ with respect to different values of λ and m. In a special case where the number of services sold by providers is uniformly distributed in $[1, N]$, we have:

$$risk_1(u_1, L) \quad = \quad u_1(1 - \phi + \frac{f(L)}{u_1})\frac{1}{N-1}$$

As proven in Section 4.3, for $\lambda < 1$ and $\phi \geqslant \phi_{min} = \frac{1-\frac{\xi}{\gamma\lambda u^*}}{1+\frac{u(1-\lambda)}{u^*\lambda}}$, it is apparent that $1 - \phi + \frac{f(L)}{u_1} < 1$. Therefore with any $L > 0$, we have $risk_1(u_1, L) < risk_2(u_2, L)$ if $m(N-1) > u_1/u_2$. In other words, for $\lambda < 1$, the system with identity premium is better for the clients as the risk of being cheated is smaller in most cases with sufficiently high N (since $u_1/u_2 < 1$).

For $\lambda > 1$, as $f(L) = u_1(1-\phi)\lambda\frac{\lambda^L - 1}{\lambda - 1}$, one can verify that the inequality $risk_1(u_1, L) < risk_2(u_2, L)$ always holds provided that:

$$(\lambda^N - 1)\frac{\lambda+1}{\lambda-1} < \frac{u_2 m(N-1)}{u_1(1-\phi)}$$

Figures 6(a) and 6(b) show the relation between the relative cheating gain ratio λ, the maximal number of services sold by providers N with which a identity premium-based price model is a better choice for risk-sensitive clients. It is observed that for $\lambda > 1$, the identity premium-based approach may still be less risky and preferable to clients if the providers only sell a small number of services N and the whitewashing intention m is sufficiently high.

In summary, a system with identity premium is acceptable for rational clients in terms of minimizing their risks in those application scenarios with low temporary cheating gain by providers such that the system parameter $\lambda < 1$. In the case of $\lambda > 1$, the mechanism is still acceptable for clients if clients buy many services from several providers, and providers sell a small limited numbers of services, such that the service price does not get very high. In case where these numbers can be estimated, the initial service price can even be set appropriately to minimize the (dis)advantages of each provider, as claimed in Theorem 2.

For example, in an eBay-like system, cheating means that the provider (seller) may gain the whole price paid to the sold article, so $\gamma = 1$ and $\lambda = 1/((1-\varepsilon)^k - \varepsilon^k) > 1$ for any $k > 0, \varepsilon < 0.5$. This means that a completely open (i.e., with cheap identities) and decentralized version of an eBay-like system is only practical if the providers sell a small number of articles of bounded prices and in any transaction a buyer accepts the risk of being cheated by an opportunistic seller with no more items to sell after the current transaction. In another example of service provisioning where providing bad services has a lower but non-zero value to the client, it is acceptable to assume that $\gamma < 1$. Given the availability of a sufficiently accurate trust modeling mechanism with small ε and with appropriate k, it is still possible that $\lambda < 1$ and thus an identity premium-based pricing model is readily applicable and accepted by the participants.

5. IMPLEMENTATION ISSUES

In this section, we discuss possible issues related to the implementation of the provider selection protocol in Def. 3 that uses an identity premium.

5.1. Implementation of Identity Premium

Our identity premium-based incentive mechanism is actually an alternative decentralized implementation of the popular solution of using entrance cost for newcomers to prevent whitewashing behavior. The identity premium can be implemented in different ways depending on the characteristics of the marketplaces. First of all, such a premium can be given to long-staying providers by modifying the matchmaking between providers and clients s.t. well-established providers are introduced to more clients for their future transactions. This approach is feasible in case providers sell non-depleted services, i.e., one service can be used for potentially many different clients. For example, this applies when providers are professional sellers of many similar articles, and being matched with more clients means higher revenues for the providers. Secondly, in case of depletable services, pricing mechanism can be implemented directly by regulation: providers may sell services at higher price. A potential approach to be investigated further is the combination of the identity premium with the traditional use of reputation: we consider the identity premium $f(L)$ as credit points exchangeable with real money. Those credit points of a provider act as its reputation image, helping it to be selected by more clients. In the case of many providers in the system, this reputation image is much more important to the provider than real money obtained from a transaction, i.e., the provider would accept to sell services at normal price but with higher probability of being selected by future consumers.

5.2. Implementation of a Dishonesty Detector

In our system, a dishonesty detector can be implemented with a reputation-based computational trust model, which uses statistical or heuristic methods to learn the behavior of an agent (the target) from several information sources. The first source comes from the performance statistics of the target in past transactions. These statistics are possibly collected from recommendations/ratings by previous partners and from personal experience of the learning peer on the target. Other information includes intrinsic features of the target itself, e.g., frequencies of posted ratings and involved transactions, location of the raters (Cornelli *et al.*, 2002), relationships with other agents (Ashri *et al.*, 2005). The behaviors we want to learn from a computational trust model in this case is the rating behavior of an agent, i.e., whether a client truthfully reports its experience. In centralized systems where the dishonesty detector is implemented and deployed centrally and completely trusted, the verifiability of a dishonesty detection is not necessary. In decentralized systems, a verifiable dishonesty detector can be implemented with a global computational trust model, such as EigenTrust (Kamvar *et al.*, 2003) or complaint-based (Aberer and Despotovic, 2001). More generally, the verification of the dishonesty detection can be done by disclosing the relevant information a client has used in the evaluation of a rating on a provider, possibly in an easy-to-validate form such as a Proof-carrying code (Necula, 1997).

We here provide more details about some other typical examples of the computational trust models that can serve the purpose of a dishonesty detector. The beta reputation system (BRS) (Whitby *et al.*, 2005) adopts an Iterated Filtering Approach (IFA) to detect or filter out the ratings to a seller that are not among the majority. TRAVOS (Teacy *et al.*, 2005) models the trustworthiness of agents that share ratings (raters), based on whether or not the past ratings given by the agents lead the client agent to successful transactions with service providers. The personalized approach (Zhang *et al.*, 2008) allows a client to model the trustworthiness of a rater by comparing the client's ratings and the rater's ratings to commonly rated providers, as well as the rater's ratings and all other clients' ratings to same providers. Both the SALE POMDP model (Oliehoek *et al.*, 2012) and the evolutionary trust model (Jiang *et al.*, 2013) allow a client a to ask another client b about b's trust assessment on a rater, and at the same time, take into account the trustworthiness of client b. The trust model (Fang *et al.*, 2013) derived from the diffusion theory makes use of social proximity between the client and the rater to evaluate the trustworthiness of the rater. In a social network, if the rater is socially closer to the client, the client will have higher trust towards the rater.

Zhang *et al.* (2008) performed extensive experiments to compare the performance of several probabilistic approaches (BRS, TRAVOS and their own personalized approach) in detecting unfair ratings. The misclassification error bounds of these reputation-based probabilistic trust models are well lower than 0.5 even under various adaptively malicious attacks by participating raters. Other empirical experimental results (Vu and Aberer, 2011) have confirmed that other computational trust models, such as those proposed in (Xiong and Liu, 2004), are also capable of classifying unreliable ratings with a small error bound ε under various attack scenarios, and thus they can be readily used to implement a dishonesty detector.

Another accurate yet expensive approach to dishonesty detection is to monitor the provider's performance to learn its actual *past* behavior to compare with the present ratings. For example, in e-commerce applications this monitoring can be done via legal investigations on suspicious transactions. In a market of Web services, monitoring agents can periodically probe the service and measure the real performance level offered by a provider to its clients.

In summary, methods to implement the dishonesty detector and identity premium for providers are available in most practical (decentralized) service provision systems. Given these building primitives, the implementation of any identity premium and reputation-aware provider selection protocol such as the one presented in this paper is realistic and achievable.

6. CONCLUSION

In this paper we have proposed a solution to prevent whitewashing attacks and incentivize honesty in open and decentralized reputation systems where cheap identities are available. We analyze the possibility of using a computational trust model with a given capability of identifying unreliable and biased ratings to effectively eliminate malicious providers and to enforce honesty of rational ones, given that the providers can change their identities at a small cost to avoid any punishment imposed by the community. The key to create incentives for honesty in these environments is an identity premium-based price model, where well-established providers are given advantages over new ones in pricing of their services. Such an identity premium-based pricing approach can also cope with the sparsity of ratings. The fact that providers staying in the system after a number of transactions already prove its capability of high quality service provision to consumers.

Since an identity premium-based price model relies directly on the acceptance of both service clients and providers, we have also analyzed the risk imposed to the clients and the potential losses caused by the use of identity premium. We have quantified the inefficiency of the system in relation with the identity cost and the characteristics of the marketplace under study. As a result, we have also identified those application settings that make such a pricing mechanism realistic and acceptable for both service clients and providers. Given a computational trust model with reasonably low misclassification error when detecting unreliable ratings, and in the case where the temporary cheating gain by providers is small, it is proven that rational clients can select the providers and determine the service prices such that intentionally malicious providers are quickly eliminated and rational providers are motivated to cooperate in all but the last transaction, even if it is possible for them to whitewash any bad reputation.

The current work is inspired by our previous work in (Vu and Aberer, 2011). More specifically, in the previous work, identities are assumed to be costly, and the provider selection protocol in Def. 3 can simply assure that any rational provider is motivated to cooperate in all but a small number of its last transactions. The current work is the further development of the previous one by introducing identity premium for the case of cheap identities. Combining the two pieces of work together, we have proposed an effective approach to using computational trust models to enforce honesty in presence of any cost model for identity.

Our analysis is limited to the case where the identity premium is only a function of the length of stay of a provider. The main reason is that its verifiability helps us to consider the accuracy of the computational trust model being used to evaluate the rating reliability. As a future work, this analysis may be extended to find the equilibria of user strategies under different service pricing mechanisms that incorporate other factors such as the number of negative and positive ratings on a provider, and to quantify the efficiency and applicability of such mechanisms. On this direction, existing mechanisms from micro-economics and operations research may be applied, e.g., providers may sell their reputation upon quitting the system, leading to a market of reputation (Tadelis, 2002).

Another direction to explore is the distributed deployment of our model by, for example, allowing service clients to set different values for k (the number of posted cheating detection on a service provider) and to use different trust models (dishonesty detectors). We will conduct simulations to validate our model in such settings and observe the correlations between k and the accuracy of different dishonesty detectors. We will also expand our model to allow clients to provide ratings in different scales, for example a real value in the range of $[0, 1]$ or a multi-scale rating, to obtain better estimation of the deceptive behavior of providers (i.e., γ in Section 3).

Also, we will investigate whether our mechanism can resist against various sophisticated cheating strategies (i.e., attacks) of providers. For example, a malicious provider may not have any good to sell, but it creates new identities to game the system by announcing the lowest price. Our system may not be able to cope with this attack. We will also continue to experimentally verify the applicability of our mechanism in different application scenarios.

REFERENCES

Aberer, K. and Despotovic, Z. (2001). Managing trust in a peer-2-peer information system. In *Proceedings of the tenth international conference on Information and knowledge management (CIKM)*.

Aberer, K., Cudré-Mauroux, P., Datta, A., Despotovic, Z., Hauswirth, M., Punceva, M., and Schmidt, R. (2003). P-Grid: a self-organizing structured P2P system. *ACM SIGMOD Record*, **32**(3), 29–33.

Ashri, R., Ramchurn, S. D., Sabater, J., Luck, M., and Jennings, N. R. (2005). Trust evaluation through relationship analysis. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (AAMAS)*, pages 1005–1011.

Ba, S. and Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premium and buyer behavior. *MIS Quarterly*, **26**(3), pp. 243–268.

Bhattacharjee, R. and Goel, A. (2005). Avoiding ballot stuffing in ebay-like reputation systems. In *Proceedings of the ACM SIGCOMM workshop on Economics of peer-to-peer systems (P2PECON)*.

Buyya, R., Stockinger, H., Giddy, J., and Abramson, D. (2001). Economic models for management of resources in peer-to-peer and grid computing. In *Proceedings of the SPIE International Conference on Commercial Applications for High-Performance Computing*, Denver, USA.

Cornelli, F., Damiani, E., Vimercati, S. C., Paraboschi, S., and Samarati, P. (2002). Choosing reputable servents in a P2P network. In *Proceedings of the 11th International Conference on World Wide Web (WWW)*.

Despotovic, Z. and Aberer, K. (2006). P2P reputation management: Probabilistic estimation vs. social networks. *Journal of Computer Networks, Special Issue on Management in Peer-to-Peer Systems: Trust, Reputation and Security*, **50**(4), 485–500.

Douceur, J. R. (2002). The sybil attack. In *IPTPS'02*, pages 251–260.

Fang, H., Zhang, J., and Thalmann, N. M. (2013). A trust model stemmed from the diffusion theory for opinion evaluation. In *Proceedings of the 12th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Feldman, M., Papadimitriou, C., Chuang, J., and Stoica, I. (2004a). Free-riding and whitewashing in peer-to-peer systems. In *PINS '04: Proceedings of the ACM SIGCOMM workshop on Practice and theory of incentives in networked systems*, pages 228–236, New York, NY, USA. ACM.

Feldman, M., Lai, K., Stoica, I., and Chuang, J. (2004b). Robust incentive techniques for peer-to-peer networks. In *EC '04: Proceedings of the 5th ACM conference on Electronic commerce*, pages 102–111, New York, NY, USA. ACM.

Feldman, M., Papadimitriou, C. H., Chuang, J., and Stoica, I. (2006). Free-riding and whitewashing in peer-to-peer systems. *IEEE Journal on Selected Areas in Communications*, **24**(5), 1010–1019.

Friedman, E. J. and Resnick, P. (2001). The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*, **10**(2), 173–199.

Ghose, A., Ipeirotis, P. G., and Sundararajan, A. (2009). The Dimensions of Reputation in Electronic Markets. *SSRN eLibrary*.

Golbeck, J. (2006). Trust on the world wide web: A survey. *Foundations and Trends in Web Science*, **1**(2), 131–197.

Hoffman, K., Zage, D., and Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Comput. Surv.*, **42**(1), 1–31.

Huang, H., Whalley, J., and Zhang, S. (2013). Multiple Nash Equilibria in Tariff Games. *Applied Economics Letters*, **20**, 332–342.

Jiang, S., Zhang, J., and Ong, Y.-S. (2013). An evolutionary model for constructing robust trust networks. In *Proceedings of the 12th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Jøsang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, **43**(2), 618–644.

Kamvar, S. D., Schlosser, M. T., and Molina, H. G. (2003). The EigenTrust algorithm for reputation management in P2P networks. In *Proc. of WWW'03*.

Kerr, R. and Cohen, R. (2010). Trust as a tradable commodity: a foundation for safe electronic marketplaces. *Computational Intelligence*, **26**, 160–182.

König, S., Hudert, S., Eymann, T., and Paolucci, M. (2008). Trust in agent societies. chapter Towards Reputation Enhanced Electronic Negotiations for Service Oriented Computing, pages 273–291. Springer-Verlag, Berlin, Heidelberg.

Landon, S. and Smith, C. E. (1998). Quality expectations, reputation, and price. *Southern Economic Journal*, **64**(3), pp. 628–647.

Marti, S. and Garcia-Molina, H. (2003). Identity crisis: Anonymity vs. reputation in p2p systems. In *Peer-to-Peer Computing*, pages 134–141.

Miller, N., Resnick, P., and Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, **51**(9), 1359–1373.

Necula, G. C. (1997). Proof-carrying code. In *POPL '97: Proceedings of the 24th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 106–119, New York, NY, USA. ACM.

Oliehoek, F. A., Gokhale, A. A., and Zhang, J. (2012). Reasoning about advisors for seller selection in e-marketplaces via POMDPs. In *Proceedings of the the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS) Workshop on Trust in Agent Societies (TRUST)*.

Papazoglou, M. P. and Georgakopoulos, D. (2003). Service-oriented computing. *Commun. ACM*, **46**(10), 24–28.

Resnick, P. and Sami, R. (2009). Sybilproof transitive trust protocols. In *EC '09: Proceedings of the tenth ACM conference on Electronic commerce*, pages 345–354, New York, NY, USA. ACM.

Resnick, P., Zeckhauser, R., Swanson, J., and Lockwood, K. (2006). The value of reputation on ebay: A controlled experiment. *Experimental Economics*, **9**(2), 79–101.

Schoenherr, T. and Mabert, V. A. (2007). Online reverse auctions: Common myths versus evolving reality. *Business Horizons*, **50**(5), 373 – 384.

Shapiro, C. (1983). Premium for high quality products as returns to reputations. *The Quarterly Journal of Economics*, **98**(4), pp. 659–680.

Shen, W., Li, X., and Doan, A. (2005). Constraint-based entity matching. In *Proceedings of the AAAI*.

Standifird, S. S. (2001). Reputation and e-commerce: ebay auctions and the asymmetrical impact of positive and negative ratings. *Journal of Management*, **27**(3), 279 – 295.

Tadelis, S. (2002). The market for reputations as an incentive mechanism. *Journal of Political Economy*, **110**(4), 854–882.

Teacy, W. T. L., Patel, J., Jennings, N. R., and Luck, M. (2005). Coping with inaccurate reputation sources: experimental analysis of a probabilistic trust model. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (AAMAS)*.

Vu, L.-H. and Aberer, K. (2011). Effective Usage of Computational Trust Models in Rational Environments. *ACM Transactions on Autonomous and Adaptive Systems*, **6(4)**(24).

Whitby, A., Jøsang, A., and Indulska, J. (2005). Filtering out unfair ratings in Bayesian reputation systems. *The Icfain Journal of Management Research*, **4**(2), 48–64.

Xiong, L. and Liu, L. (2004). PeerTrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Trans. Knowl. Data Eng.*, **16**(7), 843–857.

Yu, H., Kaminsky, M., Gibbons, P. B., and Flaxman, A. (2006). Sybilguard: defending against sybil attacks via social networks. *SIGCOMM Comput. Commun. Rev.*, **36**(4), 267–278.

Zhang, J., Sensoy, M., and Cohen, R. (2008). A detailed comparison of probabilistic approaches for coping with unfair ratings in trust and reputation systems. In *Proceedings of the international conference on Privacy, Security and Trust*.

Zhang, J., Cohen, R., and Larson, K. (2012). Combining Trust Modeling and Mechanism Design for Promoting Honesty in E-Marketplaces. *Computational Intelligence*, **28**(4), 549–578.

APPENDIX PROOF OF THEOREM 1

Theorem 1 Assume that every client uses the protocol $\mathcal{S}_k = \langle \mathcal{R}, k \rangle$ where the dishonesty detector $\mathcal{R}$ has the misclassification errors α, β upper-bounded by $\varepsilon < 0.5$, to select a reputable provider for the next transaction.

Consider any rational provider with N services to sell. Let $u_* \leqslant u_i \leqslant u^*, i = 1, ..., N$ be the base prices of the services in the i-th transaction, as bid by the winning provider in the reverse auction. Suppose that the pricing model $P(\phi, f)$ is used by the client, it follows that:

(i) If the identity cost ξ is small, the following identity premium ensures that cooperation is always the best response strategy of the provider in *any* transaction $i = 1, ..., N-1$, for any $0 < \phi_i < 1$:

$$f(L) \quad = \quad \sum_{i=1}^{L} \lambda^{L-i}(\lambda u_i(1-\phi_i) - \xi/\gamma) \text{ for } L > 0 \tag{13}$$

(ii) For $\lambda \neq 1$, let us consider the case of no competition among providers, where the base price u_i's can be assumed as a unit cost: $u_i = 1$ and $\phi_i = \phi, i = 1, ..., N$. If the identity cost $\xi < \xi_0 = \gamma\lambda(1-\phi)$, the following identity premium function is sufficient to enforce cooperation for a provider in selling every but the last service:

$$f(L) \quad = \quad ((1-\phi)\lambda - \xi/\gamma)\frac{1-\lambda^L}{1-\lambda} \text{ for } L > 0 \tag{14}$$

With $u_i = 1, \phi_i = \phi, i = 1, ..., N$, and for $\lambda = 1$, the identity premium is:

$$f(L) \quad = \quad L(1-\phi-\xi/\gamma) \text{ for } L > 0 \tag{15}$$

(iii) Let N_h be the number of transactions that a fully cooperative (honest) provider can participate till it is mistakenly blacklisted, and let N_c be the number of bad transactions an intentionally malicious provider can benefit from defecting until eliminated from the system respectively. We have $E[N_h] > 1/\varepsilon^k$ and $E[N_c] < 1/(1-\varepsilon)^k$.

Proof. We prove (i) by considering a rational provider that has a total of N services to sell. Suppose that the provider has finished $L - 1 \geqslant 0$ transactions with the current identity and gained a utility of $U > 0$. In the current L-th transaction, the provider has an identity premium of $f(L-1)$. Let $u_* \leqslant u_L \leqslant u^*$ be the current service as proposed by this provider in the reverse auction, the provider may sell the service at an adjusted price $u_L(1-\phi) + f(L-1)$. We only consider the case $1 \leqslant L < N$, i.e., the provider still has $\Delta = N - L > 0$ services to sell after finishing the current transaction (in the last transaction there is no way to enforce a full rational provider to cooperate). Denote as $U(L, \Delta)$ the best (maximized) expected utilities a provider with an identity premium $f(L)$ may get for these remaining Δ transactions. $U(0, \Delta)$ thus corresponds to the case when the provider has no identity premium: e.g., a newly joined provider or a provider that just left and then re-joined under a new identity. Note that it costs the provider an amount ξ to create an identity.

Let $0 \leqslant t, a, b, i \leqslant 1$, where $t + a + b + i = 1$, respectively be the probabilities that the current client exhibits the following rating behaviors after the transaction: *trustworthy* (provides a reliable rating), *advertising* (posts a positive rating), *badmouthing* (posts a negative rating), and non-participating (not leaving any rating, or *ignorance*). Note that possible strategic rating manipulations by any raters colluding with the current provider are all considered by the above probabilities. For example, consider the case where the provider may use a fake identity to stuff a positive rating with a newer timestamp to hide its cheating in a transaction. In this case, the provider still has an additional gain v in the transaction, and the dishonesty detection is applied on the fake rating by a client exhibiting an advertising behavior, i.e., $t = i = b = 0, a = 1$.

The probabilities that an honest provider obtains a positive [resp. negative] rating after a transaction are $h^+ = t + a + i = 1 - b$ [resp. $1 - h^+$]. The honest provider is blacklisted if the genuine positive rating is not accepted by the computational trust model as reliable, with a probability β [resp. the biased negative rating is accepted as reliable with a probability α. Thus the probability that the provider will be blacklisted by a forthcoming client is: $x_b = h^+\beta + (1-h^+)\alpha = (1-b)\beta + b\alpha \leqslant \varepsilon$, since $0 \leqslant b \leqslant 1$ and $0 \leqslant \alpha \leqslant \varepsilon, 0 \leqslant \beta \leqslant \varepsilon$.

The probability that the provider is globally blacklisted after the current transaction is $x_b^k \leqslant \varepsilon^k$. This holds even in the presence of malicious or strategic manipulation of ratings by raters with different t, a, b, i, provided that the errors α, β of $\mathcal{R}$ are less than ε.

Similarly reasoning, if the provider is cheating in this transaction, the probability the provider is globally blacklisted is $y_b^k \geqslant (1-\varepsilon)^k$. Note that the above analysis holds even in the presence of malicious or strategic manipulation of ratings by providers, provided that misclassification errors α, β of $\mathcal{R}$ are less than ε. A globally blacklisted provider with more services to sell has to join the system under a new identity with a zero identity premium.

Let U_{honest} [resp. U_{cheat}] be the best expected life-time utilities of the provider if it is honest [resp. cheating] in the current transaction, it follows that:

$$\begin{aligned} U_{honest} &= U + u_L(1-\phi) + f(L-1) + (1-x_b^k)U(L,\Delta-1) + x_b^k(U(0,\Delta-1)-\xi) \\ U_{cheat} &= U + [u_L(1-\phi) + f(L-1)](1+\gamma) + (1-y_b^k)U(L,\Delta-1) + y_b^k(U(0,\Delta-1)-\xi) \\ \delta_{hc} &= U_{honest} - U_{cheat} = -[u_L(1-\phi) + f(L-1)]\gamma + (y_b^k - x_b^k)(U(L,\Delta-1) - U(0,\Delta-1) + \xi)) \\ &\geqslant -[u_L(1-\phi) + f(L-1)]\gamma + ((1-\varepsilon)^k - \varepsilon^k)(U(L,\Delta-1) - U(0,\Delta-1) + \xi) \end{aligned} \tag{16}$$

Suppose that in the next transaction the provider sells another service with an original price u' (possibly the winning price of another reverse auction). If the provider is honest in the next transaction, we have:

$$\begin{aligned} U(L,\Delta-1) &= u'(1-\phi') + f(L) + (1-x_b^k)U(L+1,\Delta-2) + x_b^k(U(0,\Delta-2)-\xi) \\ U(0,\Delta-1) &= u'(1-\phi') + (1-x_b^k)U(1,\Delta-2) + x_b^k(U(0,\Delta-2)-\xi) \\ U(L,\Delta-1) - U(0,\Delta-1) &= f(L) + (1-x_b^k)(U(L+1,\Delta-2) - U(1,\Delta-2)) \end{aligned} \tag{17}$$

Similarly, if the provider defects in the next transaction:

$$U(L,\Delta-1) - U(0,\Delta-1) = f(L) + (1-y_b^k)(U(L+1,\Delta-2) - U(1,\Delta-2)) \tag{18}$$

From Equations (17, 18), and note that $x_b^k \leqslant \varepsilon^k \leqslant (1-\varepsilon)^k \leqslant y_b^k \leqslant 1$, we have:

$$\begin{aligned} U(L,\Delta-1) - U(0,\Delta-1) &\geqslant f(L) - f(0) + \\ &\quad \min(1-y_b^k, 1-x_b^k)(U(L+1,\Delta-2) - U(1,\Delta-2)) \\ \Rightarrow U(L,\Delta-1) - U(0,\Delta-1) &\geqslant f(L) + (1-y_b^k)(U(L+1,\Delta-2) - U(1,\Delta-2)) \end{aligned} \tag{19}$$

By similar reasoning the following recurrence relations can be found for $1 \leqslant i \leqslant \Delta - 2$:

$$\begin{aligned} U(L+i,\Delta-i-1) - U(i,\Delta-i-1) &\geqslant (1-y_b^k)(U(L+i+1,\Delta-i-2) - U(i+1,\Delta-i-2)) \\ &+ f(L+i) - f(i) \end{aligned}$$

and $U(L+\Delta,0) - U(\Delta-1,0) = f(L+\Delta) - f(\Delta-1)$.

From the above recurrences[1], if follows that:

$$U(L,\Delta-1) - U(0,\Delta-1) \geqslant f(L) + \sum_{i=1}^{\Delta-1}(1-y_b^k)^i(f(L+i) - f(i)) \tag{20}$$

Let $f(L)$ be a monotonically increasing function of L, then $f(L+i) - f(i) \geqslant 0$. From Equations (16, 20), for any $1 \leqslant L < N$, where N is the total of number of services the provider wants to sell in the system, we have:

$$\delta_{hc} = U_{honest} - U_{cheat} \geqslant -[u_L(1-\phi) + f(L-1)]\gamma + ((1-\varepsilon)^k - \varepsilon^k)(f(L) + \xi) \tag{21}$$

Cooperation in this (L-th) transaction is the best response strategy of the provider whenever $\delta_{hc} \geqslant 0$, which is always held if:

$$f(L) - \frac{\gamma}{(1-\varepsilon)^k - \varepsilon^k} f(L-1) \geqslant \frac{u_L(1-\phi)\gamma - \xi}{(1-\varepsilon)^k - \varepsilon^k} \tag{22}$$

Recall that $\lambda = \frac{\gamma}{(1-\varepsilon)^k - \varepsilon^k} > 0$. By similar reasoning, cooperation is the best response strategy for the

[1] Rigorously, the probabilities x_b^k, y_b^k may be different in each equation and thus y_b^k shall be the largest one among these y_b^k. However, to simplify the notation we will ignore such differences.

provider in all transactions $1, ..., L$ iff:

$$\begin{aligned}
f(L) - \lambda f(L-1) &\geqslant u_L(1-\phi_L)\lambda - \xi/\gamma \\
f(L-1) - \lambda f(L-2) &\geqslant u_{L-1}(1-\phi_{L-1})\lambda - \xi/\gamma \\
&\dots \\
f(2) - \lambda f(1) &\geqslant u_2(1-\phi_2)\lambda - \xi/\gamma \\
f(1) &\geqslant u_1(1-\phi_1)\lambda - \xi/\gamma, \text{ where } f(0) = 0 \\
\Rightarrow f(L) &\geqslant \sum_{i=1}^{L} \lambda^{L-i}(\lambda u_i(1-\phi_i) - \xi/\gamma)
\end{aligned}$$

The following minimal identity premium function satisfies every above constraint:

$$f(L) = \sum_{i=1}^{L} \lambda^{L-i}(\lambda u_i(1-\phi_i) - \xi/\gamma) \tag{23}$$

Let ϕ^* be the largest among $\phi_i, i = 1, ..., L$ and note that $u_i \geqslant u_*, i = 1, ..., L$, we have:

$$\begin{aligned}
f(L) - f(L-1) &= \lambda u_L(1-\phi_L) + (\lambda - 1)\sum_{i=1}^{L-1} \lambda^{L-i} u_i(1-\phi_i) - \xi/\gamma\lambda^{L-1} \\
&\geqslant \lambda u_*(1-\phi^*) + (\lambda - 1)\sum_{i=1}^{L-1} \lambda^{L-i} u_*(1-\phi^*) - \xi/\gamma\lambda^{L-1} \\
&= u_*(1-\phi^*)\lambda^L - \xi/\gamma\lambda^{L-1} > 0 \\
\Leftrightarrow \xi &< \xi_0 = u_*(1-\phi^*)\gamma\lambda
\end{aligned} \tag{24}$$

One may verify that the inequality (24) holds for both cases of $\lambda \neq 1$ and $\lambda = 1$. In other words, $f(L)$ is an increasing function of L with any $\xi < \xi_0$, and thus the requirement $f(L+i) > f(i)$ in (20) is met. Therefore, with the identity premium of (13) and where $0 < \varepsilon < 0.5$, cooperation is always the best response strategy of a rational provider in any transaction after which it still has $\Delta = N - L > 0$ services to sell. That means the provider is motivated to cooperate at every transaction $L = 1, ..., N-1$ and thus (i) is proven. For the simple case with $u_i = 1, \phi_i = \phi, i = 1, ..., N$, we have (ii) follows immediately.

To prove (iii), note that after each transaction, the probability that by accident, an honest provider is globally blacklisted is $x_b^k \leqslant \varepsilon^k$. In the worst case ever, N_h is a geometric random variable with probability ε^k, hence $E[N_h] > 1/\varepsilon^k$.

By similar reasoning, the probability a malicious provider is globally blacklisted is $y_b^k \geqslant (1-\varepsilon)^k$, and thus $E[N_c] < 1/(1-\varepsilon)^k$. □