Towards Provably Secure Trust and Reputation Systems in E-Marketplaces

Reid Kerr
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
rckerr@cs.uwaterloo.ca

Robin Cohen
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
rckerr@ai.uwaterloo.ca

ABSTRACT

In this paper, we present a framework for evaluating the security of trust and reputation systems for electronic marketplaces populated with buying and selling agents. Our proposed framework offers a method for researchers to understand the security of their systems, and to provide precise guarantees of the degree of provable security that these systems offer. We demonstrate the viability of our proposed framework by presenting a specific monetary-based trust system known as Trunits, along with an analysis that shows that Trunits provides a guaranteed level of security for buyers.

1. INTRODUCTION

A common theme of work in trust and reputation in multiagent systems (e.g., [3, 4, 5]) is increasing the likelihood of selecting a trustworthy business partner. Unfortunately, this may not be enough to inspire the confidence of potential users or adopters of such systems—anything less than complete assurances of trustworthiness raises doubts. Some trust and reputation systems attempt to provide incentives for agents to be honest (e.g., [1]), but one must still ask: Under what circumstances will the incentive hold? Will the incentive always be sufficient?

If systems aren't provably secure, the potential for vulnerabilities exists, and self-interested agents will exploit these in order to maximize profit. We have identified a catalogue of vulnerabilities in trust and reputation systems, and demonstrated that these vulnerabilities are common in existing models (see [2]). This highlights the need to consider the issue of system security carefully. Motivated by work in the field of cryptography, which seeks to deliver provable protection to users, we seek methods for ensuring the provable security of trust and reputation systems. We define security in terms of a set of 'safety properties': conditions that, if proven to hold for the system, ensure participants within the system will not be harmed. In this paper, we focus on the security of buyers, since protecting buyers from cheating sellers is a predominant focus of current research.

AAMAS'07, May 14-18, 2007, Honolulu, Hawai'i, USA.

2. THE SECURITY FRAMEWORK

In our framework, we distinguish between an *agreed transaction* $t_A = (g_p, v, d_p, A_b, A_s, \ldots)$, where g_p is the good promised, v is the value (agreed price) of the good, d_p is the date/time promised, A_b is the buying agent, and A_s is the selling agent, and a *delivered transaction* $t_D = (t_A, g_d, c, d_d, \ldots)$, where t_A is the agreed transaction, g_d is the good delivered, c is the cost incurred by the seller in providing and delivering the good, and d_d is the date/time of delivery. We consider a delivered transaction t_D to be *honest* if it fulfills the seller's commitments—g_d satisfies g_p, d_d satisfies d_p, etc.—and denote it by the predicate $honest(t_D)$. For brevity, we make use of 'accessor' functions that return the value of individual transaction parameters of the same name. We term a set of transactions a *schedule*. Let T_D represent a schedule of delivered transactions. For any T_D, there is a corresponding T_A consisting of the same transactions with the delivery parameters removed.

For any coalition of sellers G, consider T_D where $t_D \in T_D \Leftrightarrow A_s(t_D) \in G$. For each transaction in the set, the sellers in G may choose to execute the transaction honestly or dishonestly. Not all schedules can actually be executed. For example, an agent that cheats repeatedly might not continue to find buyers for its products; although it might be possible to formulate a schedule that includes continued future business, such a schedule may be impossible under the trust system. Continuing the example, if trustworthiness is rated in the interval [0, 1], and an agent's score has dropped to 0, he may not be able to engage in further transactions, even though he has inventory. We define the predicate $feasible(t, T)$ to denote that a transaction t can actually be executed within the schedule T, in the system under consideration. $Feasible(T)$ denotes that every transaction in T is feasible. We do not define feasibility further, since it will be system- or market-specific.

As we will see, profitability is a key concern when considering the security of trust systems. The profit to the seller on an individual transaction is the selling price minus the cost, or $P(t_D) = v(t_D) - c(t_D)$. The profit to a coalition on the entire set of transactions is

$$P(G, T_D) = \sum_{\{t_D \in T_D \mid A_s(t_D) \in G \wedge A_b(t_D) \notin G \wedge feasible(t_D, T_D)\}} P(t)$$

The basis for our notion of buyer security is that a buyer will be secure under a trust system if

$\forall t_D \ \ honest(t_D)$

This would be termed *full buyer security*—i.e., it is impossible for a seller to cheat a buyer. Unfortunately, this property would be extremely hard to guarantee in practice. We might relax this property, however, by specifying a set of conditions or assump-

978-81-904262-7-5 (RPS)

tions under which the security guarantee would hold, and buyers would be secure.

For each assumption, there are two means to address it:

1. External: It may be possible to ensure that an assumption actually holds for the marketplace in question, or some mechanism external to the trust/reputation system can be used to guarantee the property. Here, the system will function adequately despite the presence of the assumption.
2. Internal: It may be possible to modify the system to remove the need for the assumption, yielding a more robust system. Thus, the presence of an assumption can provide important guidance for future research, allowing meaningful progress.

The goal would be to arrive at a system for which every remaining assumption can be ensured to hold in the marketplace. Clearly stating assumptions aids understanding of the security delivered. Some important levels of security can then be defined.

Rational-agent secure: While we may not be able to guarantee that every sale is executed honestly, we may be able to design the system so that it is in a seller's best interest to be honest:

$$\text{selling agents are rational} \Rightarrow \forall t_D \ \mathit{honest}(t_D)$$

More formally, denoting a coalition of selling agents as G:

$$[\forall G \ \mathit{rational}(G)] \Rightarrow \forall t_D \ \mathit{honest}(t_D)$$

Since rational sellers are profit maximizers, the property above can be restated as:

$$[\forall G \ \forall T_{D1},T_{D2} [\ P(G, T_{D1}) > P(G, T_{D2}) \Rightarrow T_{D1} \text{ is selected}]] \Rightarrow \forall t_D \ \mathit{honest}(t_D)$$

Rational single-agent secure: Ideally, a system would make the buyer secure regardless of collusion between agents. However, collusion is notoriously difficult to combat. A lower level of security might protect agents only from sellers who are acting alone:

$$[\forall G \ \mathit{rational}(G) \wedge |G| = 1] \Rightarrow \forall t_D \ \mathit{honest}(t_D)$$

Rational single-seller-only secure: As a weaker extension of single-agent security, a system might be secure when sellers can only act as sellers.

$$[\forall G \ \mathit{rational}(G) \wedge |G| = 1 \wedge \forall t_D \ A_B(t_D) \notin G] \Rightarrow \forall t_D \ \mathit{honest}(t_D)$$

Rational infinite-transaction secure: It may be difficult to prevent dishonest sales once sellers have exhausted finite inventories. A system may make it more attractive for a seller to continue to do business than to exit the market, but this may require the seller to have the capacity to engage in sales to infinity:

$$[\forall G \ \mathit{rational}(G) \wedge \forall T_D \ [\mathit{honest}(T_D) \wedge \mathit{feasible}(T_D) \Rightarrow (\ \exists t \ \mathit{honest}(T_D \cup t) \wedge \mathit{feasible}(T_D \cup t)\ )\ ]] \Rightarrow \forall t_D \ \mathit{honest}(t_D)$$

3. SECURITY ANALYSIS OF TRUNITS

The Trunits Model (introduced in [2]) is inspired by the concept of money, making use of abstract units of trust referred to as 'trunits'. Before a sale, the seller has some quantity of trunits, representing all of the trust gained from all buyers to date. For a buyer to consider buying from a seller, the seller must possess a sufficient degree of trust, i.e., must hold sufficient trunits. The required number of trunits is tied to the price of the good. After purchasing the good, the buyer will evaluate it, relative to her expectations.

- If the good met her expectations, then the seller gains some additional quantity of trunits.
- If the good did not meet her expectations, then the seller loses some quantity of trunits.

As a seller executes honest transactions, his trunit balance grows, allowing future profitable transactions. In contrast, dishonest sales curtail future transactions. This provides the fundamental incentive for honesty. The number of trunits gained is proportional to the size of the sale.

We require that the number of trunits risked be directly tied to the value of the transaction, using the formula:

$$V = r\tau$$

where V is the value (selling price) of the transaction, τ is the number of trunits, and r is the required *risk ratio*. The trunits are put into escrow with the market operator, pending completion of the transaction. Upon completion, if the buyer rates the transaction as unsatisfactory, then the seller loses the τ trunits placed in escrow. If, on the other hand, the buyer rates the transaction as satisfactory, then the τ trunits are returned to the seller, along with some additional quantity of trunits related to the value of the transaction, for a total of:

$$(1+p)\ \tau = (1+p)\ V/r$$

where p is a *premium* or *reward* of additional trust for acting in an honest manner.

As shown below, Trunits can provide a security guarantee:

$$\begin{aligned}
&\text{Trunits is in use} \wedge \\
&\text{selling agents are rational} \wedge && \text{(A)}\\
&\text{selling agents act alone} \wedge && \text{(B)}\\
&\text{selling agents may not buy} \wedge && \text{(C)}\\
&\text{selling agents can engage in infinite honest transactions} \wedge && \text{(D)}\\
&\text{buying agents are honest} \wedge && \text{(E)}\\
&\text{selling agents can reliably meet commitments if willing} \wedge && \text{(F)}\\
&\text{cost } c \text{ is a constant percentage of selling price} && \text{(G)}\\
&\quad \Rightarrow \forall t_D \ \mathit{honest}(t_D)
\end{aligned}$$

Specified more formally:

$$\begin{aligned}
&[\text{Trunits is in use} \wedge \\
&\forall G \ \forall T_{D1},T_{D2} [\ P(G, T_{D1}) > P(G, T_{D2}) \Rightarrow T_{D1} \text{ is selected}] \wedge \\
&|G| = 1 \wedge \\
&\forall t_D \ A_B(t_D) \notin G \wedge \\
&\forall T_D \ [\mathit{honest}(T_D) \wedge \mathit{feasible}(T_D) \Rightarrow \\
&\quad (\ \exists t \ \mathit{honest}(T_D \cup t) \wedge \mathit{feasible}(T_D \cup t)\ )\ ] \wedge \\
&\text{buying agents are honest} \wedge \\
&\text{selling agents can reliably meet commitments if willing} \wedge \\
&\text{cost } c \text{ is a constant percentage of selling price}] \\
&\quad \Rightarrow \forall t_D \ \mathit{honest}(t_D)
\end{aligned}$$

Proving the guarantee

Since rational sellers choose the most profitable option, our goal is to show that, for any arbitrary schedule, profit is maximized by executing each transaction in the schedule honestly. First, we consider only finite schedules. A delivered schedule can be specified by an agreed schedule T_A and a subset of transactions in T_A that are executed dishonestly (denoted by C). Consider any two delivered schedules that share the same T_A, yet differ in terms of C. We seek to show that for any non-empty $C \subseteq T_A$,

$$P(T_A, C) < P(T_A, \varnothing).$$

The profit function for Trunits requires further consideration, regarding the value of accumulated trunits. At the end of the schedule, the seller will have earned some profit, and will have

some quantity of remaining trunits, denoted $\tau_{bal}(exit)$. While cheating might increase profit earned during the schedule, it would reduce the number of leftover trunits—since trunits can be used to earn future profits, this is a reduction in value gained by the seller. To measure this value, we introduce one additional transaction that occurs after time *exit*. In this transaction, the seller uses all remaining trunits to cheat. As discussed below, this is not to suggest that this is what the seller will or should do. Instead, we use this to determine the profit he can *assuredly* gain from his trunits, setting a lower bound on their value. Thus, for every schedule, the total profit will be the sum of the profits from honest sales, from cheating sales, and from the 'final cheat' after the schedule has completed:

$$\begin{aligned} P(T_A, C) &= (1-c)r \sum_{i \in T_A \setminus C} \tau_i + r \sum_{i \in C} \tau_i + r\,\tau_{bal}(exit) \\ &= (1-c)r \sum_{i \in T_A \setminus C} \tau_i + r \sum_{i \in C} \tau_i + r\left(\tau_{init} - \sum_{i \in T_A} \tau_i + (1+p) \sum_{i \in T_A \setminus C} \tau_i\right) \end{aligned}$$

To compare profits from each schedule, we subtract the profit of the second from that of the first. After simplification:

$$P(T_A, C_1) - P(T_A, C_2) = (1-c+p)\, r \sum_{i \in C_2 \setminus C_1} \tau_i$$

Given that $(1 - c)$, p, and r must all be greater than 0, if $C_1 \subset C_2$, the profit using C_1 must be higher than that of C_2. Thus, for any finite T_A and non-empty $C \subset T_A$, $P(T_A, C) < P(T_A, \emptyset)$.

The analysis above shows that for any finite schedule, profit is maximized through honesty, but for the last 'cheating exit' transaction. Ideally, the seller will never want to make such an exit; we now relax the finite schedule constraint. Consider any arbitrary feasible schedule T_A. A rational seller will maximize profit by executing every transaction honestly, so the profit formula simplifies to:

$$P(T_A) = r(1-c+p) \sum_{i \in T_A} \tau_i + r\,\tau_{init}$$

Instead of cheating on exit, the seller might consider executing one more honest transaction t, yielding a profit of:

$$P(T_A \cup t) = r(1-c+p)\left(\sum_{i \in T_A} \tau_i + \tau_t\right) + r\,\tau_{init}$$

Since all of r, p, $(1 - c)$, and τ_t are positive, $P(T_A \cup t) > P(T_A)$, meaning that for any given schedule, it is more profitable for the seller to add honest transactions than to cheat. This implies that the seller should never cheat, but continue selling indefinitely.

Returning to our desired safety property, then,

$$\begin{aligned} &[\text{Trunits is in use} \wedge \\ &|G| = 1 \wedge \\ &\forall t_D\, A_B(t_D) \notin G \wedge \\ &\forall T_D\ [honest(T_D) \wedge feasible(T_D) \Rightarrow \\ &\qquad (\exists t\ honest(T_D \cup t) \wedge feasible(T_D \cup t)\ )\ ] \wedge \\ &\text{buying agents are honest} \wedge \\ &\text{selling agents can reliably meet their promised specifications} \wedge \\ &\text{cost } c \text{ is a constant percentage of selling price}] \\ &\qquad \Rightarrow\ \forall T_A\ [C \subseteq T_A, |C| > 0 \Rightarrow P(G, T_A, \emptyset) > P(G, T_A, C)] \end{aligned}$$

This yields:

$$\begin{aligned} &[\forall G\ \forall T_{D1}, T_{D2}\ [\ P(G, T_{D1}) > P(G, T_{D2}) \Rightarrow T_{D1} \text{ is selected}]] \wedge \\ &\forall T_A\ [C \subseteq T_A, |C| > 0 \Rightarrow P(G, T_A, \emptyset) > P(G, T_A, C)] \\ &\qquad \Rightarrow\ \forall T_A\ (T_A, \emptyset) \text{ is selected } \Rightarrow\ \forall t_D\ \ honest(t_D) \end{aligned}$$

This analysis gives us a clear picture of exactly what guarantee basic Trunits provides, and under what conditions—Trunits has achieved a provable level of security. Certain of these conditions (A, C, in the labelled list of assumptions above) might practically hold, or be made to hold in a given marketplace; others (F, G) are assumptions that might be relaxed with further analysis. Other conditions (B, D, E) provide clear directions for future work on the system. Based on this guarantee, informed decisions can be made about whether this mechanism is appropriate for a given scenario. Moreover, clear directions have been identified for future research, in overcoming these restrictions.

4. DISCUSSION AND FUTURE WORK

To be considered secure, a system with potential vulnerabilities would need to demonstrate that it is secure against an exhaustive list of such vulnerabilities (which is difficult to obtain). In contrast, proving a system to be secure according to our framework enumerates a complete set of assumptions to be addressed. It can be argued that when a predictive trust and reputation model is being used, if an agent is rational and knows that trustworthiness is being modeled, it will affect his decisions—he might be inclined to try to act honestly, to maximize future profit. In this case, a predictive model might also be viewed as a de facto incentive mechanism. This view makes our framework especially relevant for those working on predictive approaches.

In summary, the security framework presented in this paper offers a new direction for researchers in the area of trust and reputation to promote confidence in their models for real users. The Trunits model also provides a promising direction for designing electronic marketplaces in a way that offers guarantees of security to buyers.

5. REFERENCES

[1] Braynov, S. and Sandholm, T. 2002. Incentive compatible mechanism for trust revelation. In *Proceedings of the First international Joint Conference on Autonomous Agents and Multiagent Systems: Part 1* (Bologna, Italy, July 15 - 19, 2002). AAMAS '02. ACM Press, New York, NY, 310-311.

[2] Kerr, R. and Cohen, R. 2006. Modeling Trust Using Transactional, Numerical Units. In *Proceedings of the Conference on Privacy, Security and Trust* (Markham, Ontario, Canada). PST '06.

[3] Sabater, J. and Sierra, C. 2005. Review on Computational Trust and Reputation Models. *Artif. Intell. Rev.* 24, 1 (Sep. 2005), 33-60.

[4] Tran, T. and Cohen, R. 2004. Improving User Satisfaction in Agent-Based Electronic Marketplaces by Reputation Modelling and Adjustable Product Quality. In *Proceedings of the Third international Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2* (New York, New York, July 19 - 23, 2004). International Conference on Autonomous Agents. IEEE Computer Society, Washington, DC, 828-835.

[5] Yu, B. and Singh, M.P. 2002. Distributed Reputation Management for Electronic Commerce. In *Computational Intelligence*, 18(4): 535-549, 2002.