Attacks and vulnerabilities of trust and reputation models *

Jose M. Such

1 Introduction

Agents usually need to assess either the trustworthiness or the reputation of other agents in a given system. Trust and reputation are even more important in open systems, in which previously unknown parties may interact. For instance, if a buyer agent enters into an e-marketplace for the first time, it will need to choose among all of the available seller agents. As the buyer agent has no previous interactions with the seller agent, the reputation of the seller agent in the e-marketplace can play a crucial role for the buyer agent to choose a specific seller agent.

The agent community has developed a vast number of trust and reputation models [22, 26]. However, most of them suffer from some common vulnerabilities. This means that malicious agents may be able to perform attacks that exploit these vulnerabilities. Therefore, malicious agents may be able to modify the expected behavior of these models at will. As a result, these models may even become completely useless. For instance, in our previous example, a seller agent may be able to cheat the reputation model used by the buyer agent. Thus, the buyer agent may end up interacting with a malicious agent instead of what it believes a reputable agent. This has the potential to cause many damages such as money loses. Therefore, these vulnerabilities have the potential to place the whole system in jeopardy.

In this chapter, we detail some of the most important vulnerabilities of current trust and reputation models. We also detail examples of attacks that take advantage of these vulnerabilities in order to achieve strategic manipulation of trust and reputation models. Moreover, we review in this chapter works that partially/fully address these vulnerabilities, and thus, prevent possible attacks from being successful. We particularly focus on two general kinds of vulnerabilities that have received much attention from the agent community because of their fatal consequences: identity-related vulnerabilities and collusion. We firstly detail identity-related vulnerabilities and available solutions (Section 2). Secondly, we explain how reputation can be manipulated by means of collusion and how this can be partially addressed (Section 3). Then, we briefly outline other possible attacks and vulnerabilities of trust and reputation models (Section 4). Finally, we present some concluding remarks (Section 5).

2 Identity-related Vulnerabilities

Current trust and reputation models are based on the assumption that identities are long-lived, so that ratings about a particular agent from the past are related to the same agent in the future. However, when such systems are actually used in real domains this assumption is no longer valid. For instance, an agent that has a low reputation due to its cheating behavior may be really interested in changing its identity and restarting its reputation from scratch. This is

* Note this is a preliminary draft version of the chapter *Jose M. Such. Attacks and vulnerabilities of trust and reputation models. In Agreement Technologies, pp. 467-477., 2013.* Please refer to it for the last and most up to date version of this document.

what Jøsang et al. [15] called the *change of identities* problem. This problem has also been identified by other researchers under different names (e.g. *whitewashing* [4]).

The work of Kerr and Cohen [17] shows that trust and reputation models exhibit multiple vulnerabilities that can be exploited by attacks performed by cheating agents. Among these vulnerabilities, the *re-enter* vulnerability exactly matches the *change of identities* problem exposed by Jøsang et al. They propose a simple attack that takes advantage of this vulnerability: An agent opens an account (identity) in a marketplace, uses her account to cheat for a period, then abandons it to open another.

Kerr and Cohen [17] also point out the fact that entities could create new accounts (identity in the system) at will, not only after abandoning their previous identity but also holding multiple identities at once. This is known as the *sybil* attack [14]. An example of this attack could be an agent that holds multiple identities in a marketplace and attempts to sell the same product through each of them, increasing the probability of being chosen by a potential buyer.

It is worth mentioning that this is not an authenticity problem. Interactions among entities are assured[1], i.e, an agent holding an identity is sure of being able to interact with the agent that holds the other identity. However, there is nothing which could have prevented the agent behind that identity from holding another identity previously or holding multiple identities at once. For instance, let us take a buyer agent and a seller agent in an e-marketplace. The buyer has an identity in the e-marketplace under the name of *buy1* and the seller two identities in the e-marketplace *seller1* and *seller2*. Authentication in this case means that if *buy1* is interacting with *seller1*, *buy1* is sure that it is interacting with the agent it intended to. However, *buy1* has no idea that *seller1* and *seller2* are, indeed, the very same agent.

These vulnerabilities can be more or less harmful depending on the final domain of the application. However, these vulnerabilities should be, at least, considered in domains in which trust and reputation play a crucial role. For instance, in e-marketplaces these vulnerabilities can cause users being seriously damaged by losing money. This is because a seller agent could cheat on a buyer agent, e.g., a seller may not deliver the product purchased by the buyer agent. If the seller agent repeats this over a number of transactions with other buyer agents, it could gain a very bad reputation. The point is that when the seller agent gets a very bad reputation because it does not deliver purchased products, it could simply change its identity and keep on performing the same practices, causing buyer agents to lose money.

Another example can be a social network like Last.fm[2] in which users can recommend music to each other. A user who always fails to recommend good music to other users may gain a very bad reputation. If this user creates a new account in Last.fm (a new identity in Last.fm) her reputation starts from scratch, and she is able to keep on recommending bad music. Users may be really bothered with such recommendations and move to other social networks. In this case, the one seriously damaged is the social network itself by losing users.

2.1 Problem Formulation.

Such et al. [30] formulated the problem that is behind these vulnerabilities. To this aim, they used the concept of partial identity [21]: a set of attributes[3] that identify an entity in a given context. For instance, a partial identity can be a pseudonym and a number of attributes attached to it.

They also used the concept of unlinkability[21]: "Unlinkability of two or more items of interest (e.g., subjects, messages, actions, ...) from an attacker's perspective means that within the

[1]We assume that agents are running on top of a secure Agent Platform that provides authentication to the agents running on top of them, such as [28].

[2]Last.fm `http://www.last.fm`

[3]Identity attributes can describe a great range of topics [23]. For instance, entity names, biological characteristics (only for human beings), location (permanent address, geo-location at a given time), competences (diploma, skills), social characteristics (affiliation to groups, friends), and even behaviors (personality or mood).

system (comprising these and possibly other items), the attacker cannot sufficiently distinguish whether these IOIs are related or not".

Definition 1. *The* partial identity unlinkability *problem (PIUP) states the impossibility that an agent, which takes part in a system, is able to sufficiently distinguish whether two partial identities in that system are related or not.*

This problem is what causes identity-related vulnerabilities of reputation models. It is easily observed that the *change of identities* problem is an instantiation of PIUP. For instance, an agent with an identity by which she is known to have a bad reputation, acquires another identity. From then on, other agents are unable to relate the former identity to the new acquired one. Therefore, this agent starts a fresh new reputation.

Regarding multiple identities, a similar instantiation can be made, so that an entity holds several identities and has different reputations with each of them. Thus, another entity is unable to relate the different reputations that the entity has because it is unaware that all of these identities are related to each other and to the very same entity.

2.2 Existing Solutions.

There are many works that try to address the identity-related vulnerabilities of trust and reputation models. We now describe some of them based on the approaches that they follow:

Based on Identity Infrastructures: A possible solution for these vulnerabilities is the use of *once-in-a-lifetime* partial identities [10]. A model for agent identity management based on this has been proposed in [30] and has been integrated into an agent platform as described in [31]. This model considers two kinds of partial identities: permanent and regular. Agents can only hold one permanent partial identity in a given system. Regular partial identities do not pose any limitation. Although both kinds of partial identities enable trust and reputation, only permanent partial identities guarantee that identity-related vulnerabilities are avoided. Then, agents that want to avoid identity-related vulnerabilities will only consider reputation when it is attached to a permanent partial identity. This model needs the existence of trusted third parties called Identity Providers to issue and verify partial identities. This may not be a difficulty in networks such as the Internet. However, this may not be appropriate in environments with very scarce resources such as sensor networks in which an identity infrastructure cannot be assumed.

Based on Cost: When an identity infrastructure cannot be assumed, there are other approaches such as to add a monetary cost for entering a given system [10]. Thus, a potentially malicious agent would have a sufficient incentive (if the fee is high enough compared to the benefit expected) not to re-entry the system with a new identity. The main problem of this approach is that if the cost for entering the particular system is too high, even potentially benevolent agents may choose not to enter the system because of the high cost associated to it.

Based on Social Networks: There are also other solutions for identity-related vulnerabilities of trust and reputation models that can be used when trusted third parties (such as an identity infrastructure or an entity that imposes monetary costs for entering a system) cannot be assumed [13]. Yu et al. [36] present an approach based on social networks represented as a graph in which nodes represent pseudonyms and edges represent human-established trust relationships among them in the real world. They claim that malicious users can create many pseudonyms but few trust relationships. They exploit this property to bound the number of pseudonyms to be considered for trust and reputation. However, this approach is not appropriate for open Multiagent Systems in which agents act on behalf of principals that may not be known in the real world.

Based on Mathematical Properties: There is another approach that consists of reputation models specifically designed to meet some mathematical properties that are proved to avoid identity-related vulnerabilities. For instance, Cheng et al. [7] have demonstrated several conditions using graph theory that must be satisfied when calculating reputation in order for reputation models to be resilient to sybil attacks. The only drawback of this kind of approaches is that they usually need a particular and specific way to calculate reputation ratings about an individual. Thus, this approach cannot be applied to reputation models that follow other approaches for managing reputation ratings.

3 Collusion

Collusion means that a group of agents coordinate themselves to finally achieve the manipulation of either their reputation or the reputation of other agents from outside these group. Therefore, colluding agents are able to change reputation ratings at will based on attacks that exploit this vulnerability. There are two attacks that base on collusion: *ballot stuffing* and *bad mouthing* [15, 4]. These attacks mainly differ in the final objective of manipulating the reputation model. The first one attempts that the target agent gains a good reputation while the second one attempts that the target agent gains a bad reputation. They achieve this by means of providing false positive/negative ratings about the target agent. These two attacks are now described with further detail.

In ballot stuffing, a number of agents agree on spreading positive ratings about an specific agent. Thus, this specific agent may quickly gain a very good reputation without deserving it. For instance, a number of buyer agents in an e-marketplace may be spreading positive ratings about fictitious transactions with a seller agent. Thus, this seller agent may gain a very good reputation. As a result, this seller agent can cheat other buyer agents that choose it because of its good reputation.

In bad mouthing, a number of agents agree on spreading negative ratings about an specific agent, which is the victim in this case. Therefore, a reputable agent may quickly gain a bad reputation without deserving it. For instance, a number of buyer agents in an e-marketplace may be spreading negative not-real ratings about a seller agent. Thus, this seller agent may gain a very bad reputation so that other buyer agents will not be willing to interact with this seller agent.

3.1 Existing Solutions.

There are some existing solutions to avoid collusion. All of these solutions try to avoid ballot stuffing and bad mouthing based on different approaches:

Based on Discounting Unfair Ratings: One of the approaches to avoid collusion is the discount of presumable unfair ratings. There are two main approaches to do this. According to Jøsang et al. [15] there are approaches that provide what they call *endogenous* discounting of unfair ratings and others that provide what they call *exogenous* discounting of unfair ratings.

Endogenous approaches attempt to identify unfair ratings by considering the statistical properties of the reported ratings. This is why they are called endogenous, because they identify unfair ratings based on analyzing and comparing the rating values themselves. For instance, Dellarocas [8] presents an approach based on clustering that divides ratings into fair ratings and unfair ratings, Whitby et al. [34] proposes a statistical filtering algorithm for excluding unfair ratings, and Chen & Singh [6] propose the use of collaborative filtering for grouping raters according to the ratings they give to the same objects. Although all of these approaches provide quite accurate results, they usually assume that unfair ratings are in a minority. If this assumption does not hold, these approaches are less effective and even counterproductive [34].

Exogenous approaches attempt to identify unfair ratings by considering other information such as the reputation of the agent that provides the rating and the relationship of the rating agent to the rated agent. For instance, Buchegger & Le Boudec [3] present an approach for classifying raters as trustworthy and not trustworthy based on a Bayesian reputation engine and a deviation test. Yu and Singh [35] propose a variant of the Weighted Majority Algorithm [19] to determine the weights given to each rater. Teacy et al. [32] present TRAVOS, a trust and reputation model. This model considers an initially conservative estimate of the reputation accuracy. Through repeated interactions with individual raters, this model learns to distinguish reliable from unreliable raters.

Based on Anonymity: Another possible approach is to use *controlled anonymity* [8]. This approach is based on the anonymity of buyer agents and seller agents. This can potentially minimize bad mouthing because it could be very difficult (if not impossible) for colluding agents to identify the victim. However, this may not be enough to avoid ballot stuffing. This is because the seller agent can still be able to give some hidden indications of its identity to its colluding agents. For instance, the seller agent might signal its colluding agents by pricing its products at a price having a specific decimal point.

Based on Monetary Incentives: There is another approach to avoid collusion that is based on monetary incentives. In particular, monetary incentives are given to agents so that they find more profit providing real ratings rather than providing unfair ratings. For instance, the reputation model presented by Rassmusson & Janson [24] uses incentives to ensure that paid agents tell the truth when providing ratings. A similar mechanism is proposed by Jurca et al. [16] for discouraging collusion among the agents that spread ratings. They focus on payment schemes for ratings that makes the strategy of not colluding and providing true ratings rational. Therefore, agents cannot spread false rating without suffering monetary losses. Other very similar approaches have been provided in the existing literature. For instance, the authors of [2, 18] focus on discouraging ballot stuffing by means of transaction costs (e.g. commissions) that are larger than the expected gain from colluding.

4 Other Attacks and Vulnerabilities

4.1 Discrimination

Discrimination means that an agent provides services with a given quality to one group of agents, and services with another quality to another group of agents [14]. Discrimination can be either positive or negative [8, 9]. On the one hand, negative discrimination is when an agent provides high quality services to almost every other agent except a few specific agents that it does not like. The problem is that if the number of agents being discriminated upon is relatively small, the reputation of the seller will remain good so that this agent is known to provide high quality services. On the other hand, positive discrimination is when an agent provides exceptionally high quality services to only a few agents and average services to the rest of the agents. If the number of buyers being favored is sufficiently large, their high ratings will inflate the reputation of the agent. Note that discrimination is different from unfair ratings because raters are providing their true/real/fair ratings about an agent. The point is that this agent behaves differently based on the specific agent it interacts with. However, some of the solutions that are used for preventing collusion can also be applied to avoid discrimination. For instance, the controlled anonymity and the cluster filtering approaches presented by Dellarocas [8] can be used to avoid negative discrimination and positive discrimination respectively.

4.2 Value Imbalance.

In e-commerce environments, ratings do not usually reflect the value of the transaction that is being rated. This is what is known as *value imbalance* [14, 17]. An attack that can exploit this is the following. A seller agent in an e-marketplace can perform honestly on small sales. Thus, this seller agent can get a very good reputation on that e-marketplace at a very low cost. Then, this seller agent could use the reputation gained to cheat on large sales and significantly increase its benefits. Kerr & Cohen [18] present a trust and reputation model called Commodity Trunits. This model avoids the exploitation of value imbalance because it explicitly considers the value of transactions.

4.3 Reputation Lag.

There is usually a time lag between a sale and the corresponding rating's effect on the agent's reputation [14, 17]. A seller agent could potentially exploit this vulnerability by providing a large number of low quality sales over a short period just before suffering the expected reputation degradation. Commodity Trunits [18] provides an approach to solve this based on limiting the rate at which transactions can occur.

4.4 Privacy.

Enhancing privacy is by itself of crucial importance in computer applications [29]. Moreover, for the case of applications in which trust and reputation are fundamental, privacy is required in order for the raters to provide honest ratings on sensitive topics [4]. If not, this could be the cause of some well-known problems. For instance, the eBay reputation system is not anonymous (i.e., the rater's identity is known) which leads to an average 99% of positive ratings [25]. This could be due to the fact that entities in eBay do not negatively rate other entities for fear of retaliations which could damage their own reputation and welfare.

Pavlov et al. [20] introduce several privacy-preserving schemes for computing reputation in a distributed scenario. They focus on reputation systems in which the reputation computation is very simple (e.g. the summation of reputation scores). Following a similar approach, Gudes et al. [12] propose several methods for computing the trust and reputation while preserving privacy. In this case, they propose three different methods to carry out the computations of the Knots model [11]. Two of them make use of a third party while the third one is based on one of the schemes proposed by Pavlov et al. [20]. Both approaches (that of Pavlov et al. and that of Gudes et al.) present works which are only suitable for a reduced subset of trust and reputation models because they assume a particular way of calculating trust and reputation scores.

There are also some works that focus on enhancing privacy in centralized reputation systems [33, 1, 27]. In these systems, information about the performance of a given participant is collected by a central authority which derives a reputation score for every participant, and makes all scores publicly available. These works focus on providing raters with anonymity. They do not modify the computation of reputation measures but the protocols followed to carry out the computations. These protocols are based on anonymous payment systems (such as [5]).

5 Conclusion

Over the course of this chapter, some of the most important vulnerabilities of current trust and reputation models as well as the existing works with different approaches to solve them have been detailed. While some of the works presented offer solutions to some of the aforementioned vulnerabilities which are suitable under certain conditions, further research is still needed in order to completely address them. For instance, the problem of identity-related vulnerabilities in environments in which an identity infrastructure cannot be assumed remains open.

References

[1] E. Androulaki, S. G. Choi, S. M. Bellovin, and T. Malkin. Reputation systems for anonymous networks. In *PETS '08: Proceedings of the 8th international symposium on Privacy Enhancing Technologies*, pages 202–218, Berlin, Heidelberg, 2008. Springer-Verlag.

[2] R. Bhattacharjee and A. Goel. Avoiding ballot stuffing in ebay-like reputation systems. In *Proceedings of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems*, P2PECON '05, pages 133–137, New York, NY, USA, 2005. ACM.

[3] S. Buchegger and J.-Y. L. Boudec. A robust reputation system for mobile ad-hoc networks. Technical Report IC/2003/50, EPFL-IC-LCA, 2003.

[4] E. Carrara and G. Hogben. Reputation-based systems: a security analysis. ENISA Position Paper, 2007.

[5] D. Chaum, A. Fiat, and M. Naor. Untraceable electronic cash. In *CRYPTO '88: Proceedings on Advances in cryptology*, pages 319–327, New York, NY, USA, 1990. Springer-Verlag New York, Inc.

[6] M. Chen and J. P. Singh. Computing and using reputations for internet ratings. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, EC '01, pages 154–162, New York, NY, USA, 2001. ACM.

[7] A. Cheng and E. Friedman. Sybilproof reputation mechanisms. In *Proceedings of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems*, P2PECON '05, pages 128–132, New York, NY, USA, 2005. ACM.

[8] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the 2nd ACM conference on Electronic commerce*, EC '00, pages 150–157, New York, NY, USA, 2000. ACM.

[9] M. Fasli. *Agent Technology For E-Commerce*. John Wiley & Sons, 2007.

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

[11] N. Gal-Oz, E. Gudes, and D. Hendler. A robust and knots-aware trust-based reputation model. In *Proceedings of the 2nd Joint iTrust and PST Conferences on Privacy, Trust Management and Security (IFIPTM'08)*, pages 167–182, 2008.

[12] E. Gudes, N. Gal-Oz, and A. Grubshtein. Methods for computing trust and reputation while preserving privacy. In *Proceedings of the 23rd Annual IFIP WG 11.3 Working Conference on Data and Applications Security XXIII*, pages 291–298, Berlin, Heidelberg, 2009. Springer-Verlag.

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

[14] A. Jøsang and J. Golbeck. Challenges for Robust Trust and Reputation Systems. In *Proceedings of the 5th International Workshop on Security and Trust Management (STM)*, pages 1–12, 2009.

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

[16] R. Jurca and B. Faltings. Collusion-resistant, incentive-compatible feedback payments. In *Proceedings of the 8th ACM conference on Electronic commerce*, EC '07, pages 200–209, New York, NY, USA, 2007. ACM.

[17] R. Kerr and R. Cohen. Smart cheaters do prosper: defeating trust and reputation systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 993–1000. IFAAMAS, 2009.

[18] R. Kerr and R. Cohen. Trust as a tradable commodity: A foundation for safe electronic marketplaces. *Computational Intelligence*, 26(2):160–182, 2010.

[19] N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[20] E. Pavlov, J. S. Rosenschein, and Z. Topol. Supporting privacy in decentralized additive reputation systems. In *iTrust*, pages 108–119, 2004.

[21] A. Pfitzmann and M. Hansen. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml, Aug. 2010. v0.34.

[22] I. Pinyol and J. Sabater-Mir. Computational trust and reputation models for open multiagent systems: a review. *Artificial Intelligence Review*, pages In press. DOI 10.1007/s10462–011–9277–z, 2011.

[23] K. Rannenberg, D. Royer, and A. Deuker, editors. *The Future of Identity in the Information Society: Challenges and Opportunities*. Springer Publishing Company, Incorporated, 2009.

[24] L. Rasmusson and S. Jansson. Simulated social control for secure internet commerce. In *NSPW '96: Proceedings of the 1996 workshop on New security paradigms*, pages 18–25, New York, NY, USA, 1996. ACM.

[25] P. Resnick and R. Zeckhauser. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In M. R. Baye, editor, *The Economics of the Internet and E-Commerce*, volume 11 of *Advances in Applied Microeconomics*, pages 127–157. Elsevier Science, 2002.

[26] J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24:33–60, 2005.

[27] S. Schiffner, S. Clauβ, and S. Steinbrecher. Privacy and liveness for reputation systems. In *EuroPKI*, 2009.

[28] J. M. Such, J. M. Alberola, A. Espinosa, and A. García-Fornes. A Group-oriented Secure Multiagent Platform. *Software: Practice and Experience*, 41(11):1289–1302, 2011.

[29] J. M. Such, A. Espinosa, and A. García-Fornes. A survey of privacy in multi-agent systems. *Knowledge Engineering Review*, page In press., 2012.

[30] J. M. Such, A. Espinosa, A. García-Fornes, and V. Botti. Partial identities as a foundation for trust and reputation. *Engineering Applications of Artificial Intelligence*, 24(7):1128–1136, 2011.

[31] J. M. Such, A. García-Fornes, A. Espinosa, and J. Bellver. Magentix2: a privacy-enhancing agent platform. *Engineering Applications of Artificial Intelligence*, page In press., 2012.

[32] W. Teacy, J. Patel, N. Jennings, and M. Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

[33] M. Voss. Privacy preserving online reputation systems. In *International Information Security Workshops*, pages 245–260, 2004.

[34] A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proceedings of the 7th International Workshop on Trust in Agent Societies*, New York, NY, USA, 2004.

[35] B. Yu and M. P. Singh. Detecting deception in reputation management. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, AAMAS '03, pages 73–80, New York, NY, USA, 2003. ACM.

[36] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. Sybilguard: defending against sybil attacks via social networks. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM)*, pages 267–278, New York, NY, USA, 2006. ACM.