RepHi: A Novel Attack against P2P Reputation Systems

Jingyu Feng[a,b] , Yuqing Zhang[b]*, Shenglong Chen[b], Anmin Fu[a,b]
[a] Key Lab of Computer Networks and Information Security of Ministry of Education, Xidian University,
Xi'an 710071, China
[b] National Computer Network Intrusion Protection Center, Graduate University of Chinese Academy of Sciences,
Beijing 100049, China

***Abstract*—Reputation systems are having increasing influence on guarding P2P networks. However, reputation systems themselves are vulnerable to attack. Due to the nature of aggregating ratings, the reputation scores of certain peers can be manipulated intentionally by adversaries. In this paper, we report the discovery of a novel attack, named RepHi (Reputation Hijack), against rating-based reputation systems, such as those used in P2P networks. In RepHi, attackers disguise as routers to hijack and modify ratings. This attack can cause multi-dimensional damage, that is, undermining reputation systems, manipulating reputation and hurting the credibility of raters. We conduct an investigation on RepHi, including basic ideas and case studies. Compared with other known attacks, the RepHi attackers require less efforts to achieve the similar goals.**

***Index Terms*—P2P, trust, reputation, hijack, security**

I. Introduction

The anonymous and open nature of peer-to-peer (P2P) networks offer an ideal environment to spread inauthentic files among peers, such as spread the VBS.Gnutella in Gnutella and W32.Supova.Worm in KaZaa [1]. To guard P2P networks and combat malicious peer behaviors, trust management is essential for peers to evaluate the reputation of others. A reputation system calculates the reputation score of a peer by aggregating the ratings from all raters who have interacted with this peer. By making the reputation scores publicly available, peers are able to make informed decisions about which peers to trust [2].

Currently, a number of research groups are working on applying reputation systems [3-7] to guard P2P networks, but only few of them consider seriously the vulnerabilities of reputation systems themselves. Since reputation systems highly depend on the aggregated ratings, they are vulnerable to false ratings. Examples include denial-of-service (i.e. making the system unavailable), self-promoting (i.e. falsely increasing the reputation of certain malicious peers) and slandering (i.e. falsely reducing the reputation of some honest peers) [8]. These existing attacks share one common property: forging raw data, that is, a good deal of raters are hired by attackers to submit false ratings.

However, this philosophy has a big limitation. To launch a successful attack, attackers need to consume such a mass of resources to subvert reputation systems. The resources can be the number of raters under attackers control and the maximum number of ratings can be inserted by these raters [9]. In most reputation systems, each rater can only submit one rating in a given action when receiving the query related to the evaluated peer. Additionally, when false ratings overweight actual ratings and become majority, the reputation of a peer would be manipulated easily [10]. Therefore, it is a huge task for attackers to lead a sufficient number of raters to achieve their goal.

Note that P2P protocols follow message forwarding mechanisms, where a message reaches the desired peer after going through a number of other peers in the network, hence anybody can fake critical un-encrypted messages [11]. In this paper, we discover a novel attack along this line, named as Reputation Hijack or RepHi in short, against reputation systems, and conduct an investigation on this new attack. Concretely, this paper makes three unique contributions:

- Analyze key factors that facilitate RepHi. These factors include unfamiliarity, un-encrypted ratings, and information transparency in rating aggregation.
- Introduce an attack model that inspires the discovery of RepHi. A simple idea of the attack model is as follows. In rating aggregation, peer *i* wants to know the reputation truth about peer *j* and sends a query to the network. Attackers first hijack the query, and then modify *i*'s ID such that they can receive the ratings submitted by raters. After receiving this messages, attackers tamper with and forward them back to *i*. As a result, since *i* wrongly believes the tampered ratings come from raters, attackers are able to successfully manipulate *j*' reputation. Meanwhile, the ratings from the honest peers, which disagree with *j*'s actual behavior after performing an action, are considered as dishonest, and the credibility of the honest raters is reduced.
- Difficult to defend against RepHi attack. A straightforward solution of avoiding this risk is to authenticate each rating using cryptography schemes. Unfortunately, this solution faces two performance obstacles: high transmission and computational overhead. To begin with, the size

This work is supported in part by National Science Foundation of China under Grant No. 60970140, No.60773135 and No.90718007.

*Corresponding author. E-mail address: zhangyq@gucas.ac.cn

978-1-4244-9920-5/11/$26.00 ©2011 IEEE

of digital signatures is typically very large, which will introduce extra transmission overhead. Furthermore, public key signature verification is typically computationally extensive operations, and thus verifying those individual signatures one by one at each rating will significantly increase the computational overhead. For this reason, it is difficult to apply such way to authenticate ratings in highly distributed environments like P2P networks.

The rest of the paper is organized as follows. In Section II, related work is briefly introduced. Section III outlines RepHi attack, including basic ideas, attack model and case studies. Some additional discussions are provided in Section IV. Finally, we conclude in Section V.

II. Related Work

The application of reputation systems is more effective, such as helping a peer to find out which is the most trustworthy or reputable participant to have an interaction with, preventing thus the selection of a fraudulent or malicious one [12]. However, the manipulation of such systems is rapidly growing, and several attacks have been proposed.

Whitewashing [13] is an initial attack against reputation systems, in which attackers reenter the network with a new identity to repair their reputation. Hoffman et al. [8] argue that mitigating whitewashing requires reputation systems to use a formulation that does not result in the same reputation for both newcomers and old peers, and limit peers from quickly switching identities or obtaining multiple identities. They also introduce three severe attacks: denial-of-service, self-promoting and slandering.

In denial-of-service, attackers prevent the calculation and dissemination of reputation scores. This attack is typically conducted to subvert centralized reputation systems and cause the central entity to become overloaded. Nevertheless, distributed systems are often robust if enough redundancy is employed such that loss of a few peers will not affect the calculation of reputation scores. Without the central entity, denial-of-service is difficult to be exploited in P2P networks. In fact, the basic goal of malicious attacks against a reputation system is to boost or reduce the reputation of certain peers [14]. To achieve the goal, attackers can 1) provide positive ratings for self-promoting to make malicious peers like trustworthy; 2) fake negative ratings for slandering to make their competitors look suspicious. Furthermore, both the existing attacks can be implemented by two strategies: individual (i.e. an individual rater provides false ratings) or collaborative (i.e. a group of attackers form a collusive clique to manipulate the reputation of certain peers intentionally). Compared with collaborative strategy, individual strategy is less harmful and can be addressed easily [10]. Therefore, self-promoting and slandering are usually organized in groups of collaborating identities. This two kinds of attackers fake ratings to achieve their goal. For simplicity, this paper refers them as RawFake.

Sybil attack [15] is also exploited by attacks to combine with self-promoting and slandering, in which attackers apply for a large number of identities to submit false ratings. Fortunately, such attack can be defended easily, for two reasons. 1) Many systems limit the number of user IDs per IP address [14]. 2) Some techniques [16] can be utilized to increase the cost for a malicious server to control many user IDs greatly. In addition, RepTrap [14] is presented to attack reputation systems with limited resources, in which attackers hurt honest raters' credibility and improve their own credibility at the same time. However, such attack is also employed to centralized reputation systems. It is important to point out that many applications provide services in a distributed manner, such as P2P networks.

Overall, in these attacks, more attention has been paid to RawFake recently. However, to launch a successful RawFake, attackers need to form a collusive clique and become the majority of raters. In this paper, we discover RepHi, which is difficult to defend and demands less efforts than RawFake.

III. RepHi Attack

In this section, we describe the basic ideas of RepHi and present an attack model. Afterward, we illustrate this attack with case studies.

A. Basic Ideas

In the area of file sharing in P2P networks, each peer plays two roles, the role of server providing files to other peers and the role of consumer sharing files. It is important to establish trust between two parties for a specific action, especially when the consumer is not sure about the reputation of the server. The attack goal against reputation systems is to mistake the judgement of a consumer with less resources. This inspires us to discover RepHi attack.

In RepHi, attackers hijack some ratings reported to a server, and then modify them to mislead a possible consumer. That is, the consumer makes wrong judgement about the server reputation through aggregating false ratings. Furthermore, the consumer thinks the ratings from raters disagree with the server's behaviors. Therefore, the reputation of the server will be manipulated while the credibility of raters will be reduced. The common practice is to assume that the ratings reported by raters are actual.

RepHi is applicable under three key factors. 1) Participants are not familiar with each other. In P2P networks, a consumer wants to gain resources from a server, but it worries that the server might be vicious or malicious. 2) Ratings are unencrypted. None of effect cryptography technologies have been considered to encrypt and authenticate ratings in current reputation systems. 3) Query and ratings are transparent between a consumer and raters. The topology of P2P networks graph is meshed, and all peers also act as routers forwarding incoming messages to neighbors. A message reaches the desired peers after going through a number of other peers in the network. In this case, ratings can be hijacked easily.

Overall, instead of faking ratings, the RepHi attackers hurt innocent consumers at a relatively low cost. RepHi also has other effects, such as hurting the incentive of honest raters, magnifying the conspirators of attackers, or even slandering their competitors.

B. Attack Model

This section illustrates an attack model to launch the RepHi attack heuristically. In this model, we classify all the participators into four categories.

- **Target servers (*TS*s).** Servers are the peers who play the role of offering files. When some servers become the conspirators or competitors of attackers, they are *TS*s. For conspirators, attackers improve their reputation by false high ratings. If these conspirators are malicious, they would be more dangerous than normal due to their high reputation. Conversely, attackers reduce the reputation of competitors. Hence, these competitors would be suspected as malicious.
- **Target raters (*TR*s).** Raters who have interacted with *TS*s submit their opinions to consumers. Since raters may fake raw ratings, some researchers [4,5,10] evaluate the credibility of raters to filter out false ratings. More importantly, a rater's credibility is used to weigh its rating in the calculation of reputation scores. Through tampering with the ratings in the forward, the RepHi attackers can undermine the credibility of *TR*s for their purpose.
- **Innocent consumers (*IC*s).** Consumers are the peers who play the role of sharing files. *IC*s will make wrong judgement due to the manipulated reputation of *TS*s.
- **Malicious routers (*MR*s).** In P2P networks, routers are the neighbors of peers, in charge of forwarding messages. *MR*s, called the RepHi attackers in this paper, hijack and tamper with the ratings from *TR*s to *IC*s.

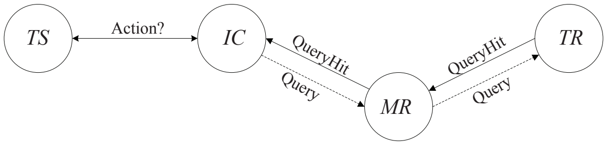

Fig. 1. Relationship among participators.

As shown in Fig.1, we construct the attack model by four procedures.

Step P1. Peer *i* finds peer *j* who holds the requested files. To ensure the security, *i* broadcasts a Query message $Q[ij] = (ID_i, ID_j, R, ttl)$ to collect the ratings about *j*. The query are labeled by a unique ID of two participators and can be used by the recipient to detect where the Query comes from. R indicates that *i* looks for the reputation of *j*.

Step P2. Peer *m* is a RepHi attacker who wants to manipulate the reputation of *j*. As a router, *m* monitors all the Query messages. When $Q[ij]$ appears, *m* becomes a *MR* and changes ID_i into ID_m. At the same time, *i* and *j* are turned into an *IC* and *TS* respectively.

Step P3. Peer *k* that receives $Q[ij]$ and has transacted with *j*, answer with a QueryHit unicast message $QH[ij] = (ID_m, ID_j, ID_k, r_{kj}, TS)$ to *m*. r_{kj} is the rating provided by *k*. TS is the current time, which is used to prevent the QueryHit message replay.

Step P4. Upon reception of $QH[ij]$, *m* modifies r_{kj} for its purpose, and changes ID_m back to ID_i. At last, *m* sends the modified $QH[ij]$ to *i*.

C. Case Studies

To further illustrate RepHi, we create an scenario in a basic reputation system. In this subsection, we firstly describe this basic reputation system, then present the scenario.

1) A Basic Reputation System: For reputation systems, one of the most popular designs is based on averaging rating. Liang and Shi [17] have proven that the simple averaging rating is good enough considering the simplicity of the algorithm design, and the low cost in the system running. On this basis, we abstract a basic reputation system with averaging rating to demonstrate the RepHi attack.

The basic reputation system first evaluates the credibility of raters, and then calculates reputation scores based on averaging rating. If we combine them together, the system is as follows.

- **Rater credibility algorithm:** This algorithm that evaluates the credibility of raters to weigh their ratings is referred to as *RC Algorithm*. Let $\overline{r}$ be the average of all the ratings. r_{kj} is the rating given to *j* by *k*. The credibility of *k*, denoted by C_k, is calculated as the bias between r_{kj} and $\overline{r}$.

$$C_k = 1 - |r_{kj} - \overline{r}| \tag{1}$$

  Where, $\overline{r}$ represents the majority opinion of all the raters. It is worth noting that the majority rule is one of the most popular designs in reputation systems [9]. Briefly speaking, if *k*'s rating disagrees (or agrees) with the majority opinion, the credibility of *k* would be reduced (or increased).
- **Server reputation algorithm:** This algorithm that calculates reputation scores is referred to as *SR Algorithm*. A server's reputation score is calculated based on ratings and the *RC Algorithm*. Let RS_j denote *j*'s reputation score and Φ is the set of raters. j's reputation score is calculated as:

$$RS_j = \sum_{k \in \Phi} r_{kj} * \frac{C_k}{\sum_k C_k} \tag{2}$$

The key to the robustness of the basic reputation system is the *RC Algorithm*. This is because reputation systems are vulnerable to false ratings. In the *RC Algorithm*, only when false ratings become the majority opinion can attackers get high credibility. Consequently, attackers can manipulate a server's reputation in the *SR Algorithm*.

2) A Simple Scenario: To illustrate the basic idea of RepHi and compare with RawFake, we present a detailed example in this simple scenario.

Assume that *j* is marked as a malicious server for $RS_j = 0.3$. If all the ratings are real, their value should range from

0.2 to 0.4. Let the threshold of reputation score is 0.5. In this case study, attackers' goal is to make *j* as trustworthy for $RS_j > 0.5$.

Under the constraint that one rater only summits one rating, the attack effort depends on two factors (α, β), the number of raters and the percentage of dishonest raters respectively. In this scenario, we perform a experiment over the basic reputation system to observe the manipulation of *j*'s reputation by the two factors.

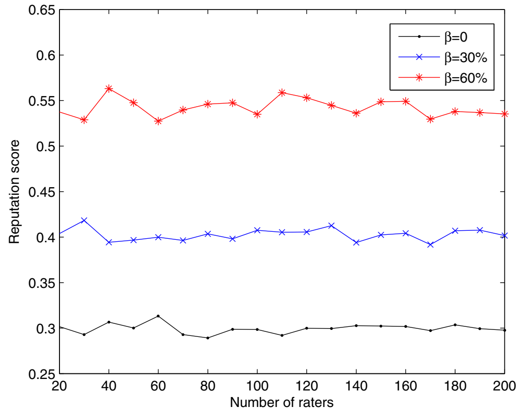

Fig. 2. Manipulating *j*'s reputation using two factors.

As shown in Fig.2, we can find an interesting observation: the attack effort mainly relies on the percentage of dishonest raters, rather than the number of raters. The higher is β, the higher reputation *j* will gain. Specially, *j* begins to get large reputation score when β is more than 60%. In this sense, *i* may trust *j* for its high reputation. As a consequence, the goal of attackers is then translated into controlling a certain number of raters to provide false ratings. Concretely, 0.6α can be referred to as *attack value*.

RawFake: In RawFake, attackers collude with one other to form a clique. That is, these attackers are dishonest raters and report false ratings. As $\alpha = 50$, the *attack value* is 30. Specifically, at least 30 attackers are required to successfully launch RawFake.

RepHi: In RepHi, attackers disguise as *MR*s. Since these attackers work on the forwarding of ratings, four parameters related to P2P networks should be taken into account.

- **Neighbors** (*nei*)**.** In P2P networks, each peer is linked dynamically to a small number of neighbors, usually between 2 and 12 [18]. So, a consumer broadcasts the Query messages to all its neighbors, then each neighbor does so until reaching the destination.
- **Time-to-live** (*ttl*)**.** To reduce network congestion, all the messages exchanged on the network are characterized by a given ttl. On passing through a peer, the ttl of a Query message is decreased by one; when the ttl reaches zero, the message is dropped. In Fig.3, we fix nei at 3. The amount of peers forwarding the Query message (N_f) is $3 \times (1 + 2^1 + 2^2) = 21$. This conclusion can be further extended as

$$N_f = 3 * \sum_n 2^{n-1}, 1 \le n \le ttl \qquad (3)$$

.

- **Distance** (dis)**.** Upon hijacking a Query message, the message has been through a few peers for the *IC* to *MR*. We define the amount of these peers as dis. In Fig.3, we fix dis at 1. The amount of peers forwarding the modified Query message (N_{fm}) is $2^1 + 2^2 = 6$. This conclusion can be further extended as

$$N_{fm} = \sum_n 2^n, 1 \le n \le ttl - dis \qquad (4)$$

.

- **Probability** (pro)**.** This parameter denotes the probability of receiving QueryHit messages for an *MR*. Finding an optimal way to count pro is difficult for two reasons. 1)It is impossible that all of peers have transacted with *j*. 2) Raters are distributed in the network randomly. With loss of generality, pro is assigned to a minimum in this scenario, such as 0.1. Specially, for a given dis, the amount of *TR*s (N_{tr}) is $0.1 * N_{fm}$.

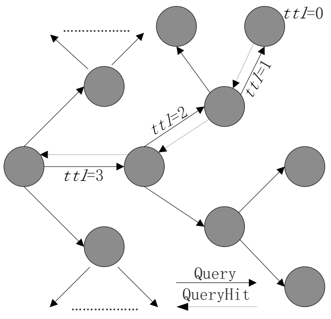

Fig. 3. Ratings collecting at $nei = 3$.

With the four parameters in mind, we can illustrate RepHi explicitly. Assume that $(nei, ttl) = (3, 8)$, a Query message can be broadcasted to 765 peers by equation (3). At $dis = 4$, an *MR* broadcasts a modified Query message to 30 peers by equation (4) and receives 3 QueryHi messages from *TR*s. Since the *attack value* is 30, at least 10 *MR*s (i.e. attackers) are needed to successfully launch RepHi.

In this case study, RepHi reduces the requirement on the number of attackers compared with RawFake. Additionally, the attack effort is proportional to nei and ttl, whereas inversely proportional to dis. Similarly, attackers can conduct RepHi in the same way to slander their competitors.

IV. Further Discussion

Reputation systems are having increasing influence on e-commerce, online digital content distribution, file-sharing, and even political campaigns. Through analysis, we have seen that RepHi can severely degrade the reputation of peers and rating credibility. Comparing with RawFake, RepHi needs the lower number of attackers to earn the same aim.

Actually, RepHi has even broader influence. We use P2P reputation systems as an example to illustrate it.

- *RepHi can promote the spread of vicious resources in P2P networks*. By manipulating the reputation of malicious peers, this attack enables malicious peers to escape from the detection mechanism of reputation systems. Therefore, malicious peers can easily spread vicious resources in P2P networks.
- *RepHi leads to the existence of inequality*. Many reputation systems reward good peers and punish malicious peers. With RepHi attack, malicious peers would not receive the punishment they deserve, or even receive the reward if they are seen as trustworthy. Contrarily, some good peers may be mistaken as vicious. This will lead to inequality, and thus affecting the enthusiasm of peers for participating P2P networks.
- *RepHi hurts the incentive to using reputation systems*. The application of reputation system is at the cost of increasing the load of networks. If reputation systems fail in predicting peers' behaviors, they would be discarded.

For future work, the following research tasks are worthy of further study.

- *Seek an effective way to defend against RepHi*. To ensure the integrity of ratings, one appealing solution is to sign each rating with a digital signature technique before the rating is sent. However, conventional signature schemes that verify the received messages one after the other may fail to satisfy the stringent time requirement of reputation systems. Note that a consumer could handle with hundreds of ratings. In this case, verifying a large number of signatures sequentially could take a long time and will certainly become the processing bottleneck at a consumer. Thus, how to minimize the security cost and improve the rating authentication efficiency becomes another critical problem for reputation systems. If a lightweight defense scheme is designed, the RepHi attack would be thwarted.
- *Research the application of RepHi in other distributed networks*. Reputation systems also play important roles in wireless sensor networks and ad-hoc networks, in which packets are also forwarded one by one. So, reputation systems in the two networks are also faced with possible RepHi attack.

V. Conclusion

Due to the nature of aggregating ratings, reputation systems themselves are an attractive target for attackers. In this paper, we discover a novel attack, named RepHi, against rating-based reputation systems. In P2P networks, the RepHi attackers disguise as routers to hijack and modify ratings. This renders these adversaries not only manipulate the reputation of a certain peer, but also hurt the credibility of *TR*s. Through an in-depth investigation, we have demonstrated that the RepHi attack can more effectively reduce the resources required as compared to RawFake. The broader impact of this attack and future work are also discussed. We are continuing our efforts in understanding potential attacks and developing robust defense schemes for securing reputation systems.

References

[1] O.H. Kwon, J. Kim, "FileTrust: reputation management for reliable resource sharing in structured peer-to-peer networks," IEICE Trans. Information and Systems, vol.E90-B, no.4, April 2007.

[2] R. F Zhou, K. Hwang, "PowerTrust: A robust and scalable reputation system for trusted P2P computing, IEEE Trans. Parallel and Distributed Systems, vol.18, no.7, pp. 460-473, April 2007.

[3] K. Aberer and Z. Despotovic, "Managing Trust in a Peer-to-Peer Information System, Proc. 10th Intl Conf. Information and Knowledge Management, 2001.

[4] S. Kamvar, M. Schlosser, and H. Garcia-Molina, "The Eigentrust Algorithm for Reputation Management in P2P Networks, Proc. ACM World Wide Web Conf. (WWW' 03), May 2003.

[5] L. Xiong and L. Liu, "PeerTrust: Supporting Reputation-Based Trust for Peer-to-Peer Electronic Communities, IEEE Trans. Knowledge and Data Eng., vol. 16, no. 7, pp. 843-857, 2004.

[6] R. Zhou, K. Hwang, and M. Cai, "GossipTrust for Fast Reputation Aggregation in Peer-to-Peer Networks, IEEE Trans. Knowledge and Data Eng., 2008.

[7] P. Dewan and P. Dasgupta, "P2P Reputation Management Using Distributed Identities and Decentralized Recommendation Chains", IEEE Trans. Knowledge and Data Eng., vol. 22, no. 7, pp. 1000-1013, 2010.

[8] K. Hoffman, D. Zage and C. Nita-Rotaru, "A Survey of Attack and Defense Techniques for Reputation System, ACM Computing Surveys, vol.41, no. 1, pp.1-31, Dec. 2009.

[9] Q. Feng, Y.L Sun, L. Liu, Y.i Yang, Y. Dai,"Voting Systems with Trust Mechanisms in Cyberspace: Vulnerabilities and Defenses," IEEE Trans. Knowledge and Data Eng., 2010.(Accepted)

[10] Y. Yang, S. Sun, S. Kay, Q. Yang, "Securing Rating Aggregation Systems using Statistical Detectors and Trust, IEEE Trans. Information Forensics and Security, vol.4, no.4, pp.883-898, Dec. 2009.

[11] E.K Lua, J. Crowcroft, M. Pias, S. Lim. "A survey and comparison of peer-to-peer overlay network schemes," IEEE Communications Surveys & Tutorials, vol.7, no.2, pp.72-93, 2005.

[12] A. Singh and L. Liu, "TrustMe: Anonymous Management of Trust Relationships in Decentralized P2P Systems, in Proc. IEEE Intl Conf. Peer-to-Peer Computing, Sept. 2003.

[13] F.G Mármol and G.M Pérez, "Security threats scenarios in trust and reputation models for distributed systems," Computers & Security, vol.28, no.7, pp.545-556, Oct. 2009.

[14] M. Feldman, C. Papadimitriou, J. Chuang, I. Stoica,"Free-Riding and Whitewashing in Peer-to-Peer Systems," in Proc. ACM SIGCOMM, Aug. 2004.

[15] Y. Yang, Q. Feng, Y. L. Sun, and Y. Dai, "Reptrap: a novel attack on feedback-based reputation systems, in Proc. 4th international conference on Security and privacy in communication netowrks, Sept. 2008.

[16] J. R. Douceur, "The Sybil attack," in Proc. the 1st International Workshop on Peer-to-Peer Systems (IPTPS). Springer, Berlin/Heidelberg, Germany, 2002.

[17] H. Yu, P. Gibbons, M. Kaminsky, and F. Xiao, "Sybillimit: A near-optimal social network defense against sybil attacks," in Proc. IEEE Symposium on Security and Privacy, May 2008.

[18] Z. Liang, W. Shi, "Analysis of ratings on trust inference in open environments," Performance Evaluation, vol.65, no.2, pp.99-128, Feb. 2008.