Robustness of Trust and Reputation Systems: Does it Matter?

Audun Jøsang

University of Oslo, Norway
josang@mn.uio.no

Abstract. Trust and reputation systems provide a foundation for security, stability, and efficiency in the online environment because of their ability to stimulate quality and to sanction poor quality. Trust and reputation scores are assumed to represent and predict future quality and behaviour and thereby to provide valuable decision support for relying parties. This assumption depends on two factors, primarily that trust and reputation scores faithfully reflect past observed quality, and secondly that future quality will be truly similar to that represented by the scores. Unfortunately, poor robustness of trust and reputation systems often makes it relatively easy to manipulate these factors, so that the fundamental assumption behind trust and reputation systems becomes questionable. On this background we discuss to what degree robustness against strategic manipulation is important for the usefulness of trust and reputation systems in general.

This paper is the printed version of the inaugural William Winsborough Commemorative Address at the IFIP Trust Management Conference 2012 in Surat.

1 Introduction

Online markets and communities are commonly moderated by trust and reputation systems, called TRS hereafter. The explosion in the use of collaborative trust and reputation propagation was triggered primarily by the speed and efficiency of the Internet and modern computers for collecting and propagating reputation information, and secondly by the emergence of Web 2.0 platforms and people's active engagement in them. Through collaborative effort members of the community provide ratings and reviews about targets which e.g. can be online services and resources as well as other community members and physical world goods and service, for example hotels, universities and medical doctors[10]. Cumulated ratings and reviews about a given target can assist other parties in deciding whether or not to use, transact with or connect with that target in the future. Introducing such systems in a community or market has multiple interrelated effects. The most direct effect is that it provides decision support for relying parties, by choosing the targets with the best scores or reviews. Targets that want to attract the business of relying parties in the future know that they need a high reputation score for that. The principle that future reputation depends on present behaviour typically influences present behaviour through the *"shadow of the future"* effect [21], meaning that anticipated future reputation casts a controlling shadow on present behaviour.

A *trust scope* refers to the specific function or quality that the target is assumed to have for the purpose of the trust relationship. In other words, the target is relied upon to have certain qualities, and the scope is what the relying party assumes those qualities to be. For example, providing financial advice and providing medical advice represent two different scopes for which trust and reputation should be considered separately. Trust and reputation also exist within a context. It should be noted that the term context is sometimes used in the sense of scope in the literature.

The term ”context” generally means the surroundings, circumstances, environment, background or settings which determine, specify, or clarify the meaning of something. We therefore define *trust context* to cover elements such as the legal and cultural environment, the domain policy, ethics and social attitudes of participants. A specific online market or social community such as eBay or facebook is always embedded in a wide context that consists of the above mentioned elements as well as others. The context of a TRS can therefore take a rather general meaning that would difficult to specify exhaustively. It would be practical to consider a domain identity such as ”eBay” or ”facebook” as an attribute of, or maybe the name/identifier of the context itself, because it indirectly refers to all its elements such as those elements mentioned above.

Another aspect of trust context is that two communities might use the same term for a specific trust scope such as ”politeness”, but the meaning could have different qualitative and semantic value if the two communities have different cultures. A simple way to convey this fact might be to include the name of the community/context as metadata or as an attribute of specific reputation scores. Another issue worth considering when comparing different TRS domains is the possibility that participants deliberately behave differently in specific different communities, so that it would not be meaningful to compute an average/federated reputation score for a specific participant who behaves in that way. In fact the community name becomes an attribute of the behaviour, i.e. the participant consciously behaves in a specific way in each different community and context. It would be possible for the participant to use the same name in the different communities so that relying parties would be aware of the difference in behaviour, or the participant could use different pseudonyms so that relying parties would ignore that two separate pseudonyms represent the same participant.

In relation to trust systems the term ”recommendation” is often used in the sense of a trust measure passed between entities, whereas the term ”rating” is often used with relation to reputation systems. In this presentation we will use the term ”rating” to denote both. The term ”score” primarily refers to a measure of trust or reputation derived by a TRS function based on the received ratings.

Many web sites allow participants to write reviews in natural language, not just as a numerical rating. For generality we will use the term ”rating” also in the sense of a review. Similarly, we let the term ”‘score” also represent the collection of reviews that are presented to the public through a website, not just a numerical trust or reputation score.

Attempts to misrepresent quality and to manipulate reputation are commonplace in human societies, and probably also in animal societies. Con artists employ methods to appear trustworthy, e.g. through skillful acting or through the fabrication and presentation of false credentials. Analogous types of attacks are being used in online communi-

ties and markets. In case of online TRSs, vulnerabilities in the systems themselves can open up additional attack vectors. From that perspective TRSs should be robust against attacks that could lead to misleading trust and reputation scores. In the worst case, a vulnerable reputation system could be turned around and used as an attack tool to maliciously manipulate the computation and dissemination of scores. The consequence of this could be a total loss of community trust caused by the inability to sanction and avoid low quality and deceptive services and agents.

Attacks against TRSs are not normally committed by computer hackers breaking into the server where the TRS functions are being hosted, although of course this could happen. Attacks against TRSs typically consist of playing the role of relying parties and/or service provider, and of manipulating the TRS through specific behaviour that is contrary to policy and/or to assumed ethical behaviour. For example, a malicious party that colludes with the service provider, or simply an unethical service provider, could provide fake or unfair positive ratings to a reputation system with the purpose of inflating the service provider's score, thereby increasing the probability of that service provider being selected by other relying parties, which in turn would lead to increased profit. Alternatively, an unethical service provider could engage in unfair badmouthing of competitors in order to reduce their business and profit, with in turn would result in increased own business and profit.

Many other attack scenarios can be imagined that, if successful, would give unfair advantages to the attackers. All such attacks have in common that that they result in the erosion of community trust, with damaging consequences for services and applications in the affected market or community. The robustness of TRSs can therefore be crucial for the quality of markets and communities where a TRS is being applied.

A TRS must not only be robust against intentional attacks, but should produce quality trust and reputation scores under changing conditions and in the presence of unsophisticated participants. Assuming that ratings provided by the community are fair one would expect that a quality service provider always is represented as such through its trust and reputation scores published through the TRS. If that is not the case, i.e. if a reliable service provider is represented with a low score and bad reviews, or an unreliable service provider is represented with a high score and good reviews, then the TRS does not fulfill its most basic role, which could be very damaging for the community. In economic terms, this could cause severe inefficiencies similarly to those resulting from corruption. A second important TRS requirement is that it must react swiftly when the rating trend changes in the positive or negative direction, by immediately producing correspondingly more positive or more negative scores [21].

A TRS can be attacked from multiple angles, meaning that designing adequate defence against possible threats can be a daunting challenge. This presentation focuses on the need for robustness in real implementations of TRS i communities and markets. We do not focus on traditional security threats such as hacking and denial of service, although such defences must of course also be included in any practical implementation. Given that each community has its own specific characteristics the need for TRS robustness will different in each case. At the same time, there are some fundamental requirements for robustness that should be satisfied in general.

2 Threat Analysis and Proposed Solutions

Fig.1 illustrates potential attack vectors related to a TRS integrated with targets and relying parties in a community or market. Note that Fig.1 represents a functional view, not an architectural view. It is for example possible that the TRS function is distributed among all the relying parties as in case of a TRS for P2P networks. It is also possible that there is no distinction between relying parties and service providers.

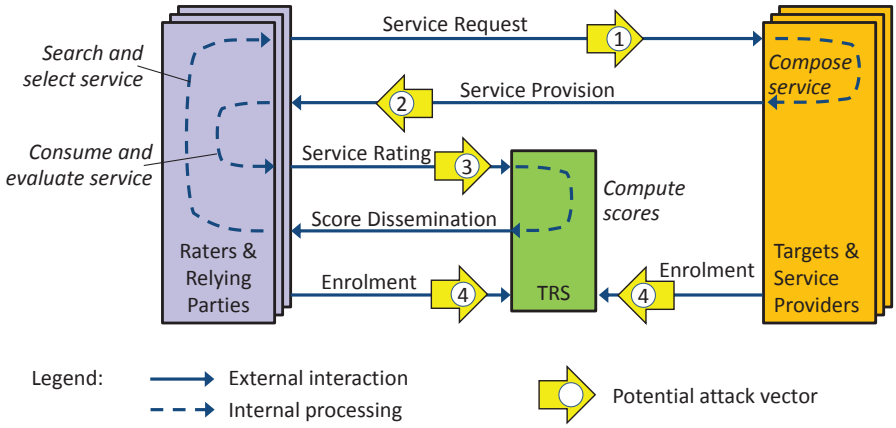

Fig. 1. Potential attack vectors related to a TRS environment

The combination of a TRS and a large number of participants (relying parties and targets) represents a highly dynamic and complex feedback system with many potential vulnerabilities. Making such systems robust against malicious manipulation represents a daunting challenge. The attack vectors in Fig.1 are briefly describe in Table 1.

Attack Vector	**Brief Description**
(1) Service Request	Malicious relying parties, possibly colluding with the service provider, could request services for the sole purpose of being entitled to rate. For example on eBay, ratings can only be provided after a registered transaction, which provides a ticket to rate.
(2) Service Provision	Malicious service providers could deliberately provide low quality services. Alternatively, low quality service could simply be the result of incompetent or unreliable service providers.
(3) Service rating	Ratings or reviews could be false or could unfairly misrepresent the actual service received.
(4) Enrollment	Relying parties and service providers can e.g. enroll multiple times in order to strategically manipulate the TRS.

Table 1. TRS attack vectors with reference to Fig1

The research literature on TRSs (Trust and Reputation Systems) is relatively mature, where the PhD thesis of Marsh (1994) [16] represents an early study of computational trust systems and the article by Resnick *et al.* (2000) [21] represents an early introduction to reputation systems. This literature is currently substantial and is still growing fast [8, 12]. A large number of TRS designs and architectures have been and continue to be proposed and implemented. Commercial implementations of TRSs are now part of mainstream Web technology which has resulted in general textbooks on how to build TRSs in real applications, such as Farmer & Glass (2011) [6].

However, the literature specifically focusing on the robustness of TRSs is much more limited and still in an early stage. It should be noted that publications on TRSs usually analyse robustness to a certain extent, but typically only consider a very limited set of attacks. The text book by Farmer & Glass [6] also offers advice on robustness. However, many studies on robustness in the research literature suffer from the authors' desire to put their own TRS designs in a positive light, with the result that the robustness analyses often are too superficial and fail to consider realistic attacks. Publications providing comprehensive robustness analyses are rare.

Hoffmann, Zage and Nita-Rotaru (2009) [9] provide a taxonomy and analysis framework for TRSs proposed for P2P networks, and then give an analysis of 24 of the most prominent TRSs based on 25 different attributes. Out of the 24 TRSs, 6 were analysed in more detail because of the representativeness of their characteristics. General challenges for the building robustness into TRSs are presented in Jøsang & Golbeck (2009) [11]. They give an overview of typical attacks described in the literature, such as those listed in Table 2.

Attack type	**Short Description**
Playbooks	Planned sequence of actions in order to manipulate and deceive
Unfair Ratings	Ratings that do not correctly reflect the actual experience
Review Spam	(aka. opinion spam) False reviews, often in conjunction with unfair ratings
Discrimination	Deliberately providing different quality services to specific relying parties
Collusion	Coordinated actions among participants in order to manipulate and deceive
Proliferation	Multiple offerings of the same service in order to obscure competing services
Reputation Lag	Abuse multiple buyers before the TRS reacts to their negative feedbacks
Re-entry	Take new identity, in order to eliminate bad reputation of old identity
Value Imbalance	Exploit reputation from many low value services, for one high value fraud
The Sybil Attack	Take on multiple identities in order to generate rating and review spam

Table 2. Various strategies for attacking trust and reputation systems

Early proposals for strengthening the robustness of TRSs were typically based on the assumption that false or unfair ratings could be detected statistically, and focused on detecting patterns and outlayers among the ratings, e.g. Dellarocas (2000) [3], Yu & Singh (2003) [26], and Withby & Jøsang (2004) [23]. While these approaches could be characterised as simplistic and relatively easy to bypass for determined attackers, they

present the idea of using data mining and reasoning to detect and protect against such attacks.

Kerr (2009) [13] provides independent robustness analyses of a set of proposed TRSs, and thereby represents a step in the right direction for TRS research. They also propose a testbed for evaluating TRSs in [14].

Large commercial TRSs have attracted, and continue to attract, the attention of independent third party analysts. For example, the robustness of Google's PageRank algorithm has been analysed by Zhang *et al.* (2004) [27] and by Clausen (2004) [2], and the robustness of eBay's Feedback Forum has been analysed by several authors, including Resnick *et al.* (2006) [22] and Dini & Spagnolo (2009) [4].

The relative simplicity of writing false reviews of goods and services, and the lack of sanctioning of this practice, currently is a significant problem and a major challenge for review sites such as epinions.com and tripadvisor.com. The problem of false reviews seems difficult to solve because it is in principle impossible to read people's minds and verify whether a review really reflects their inner thoughts. There are nevertheless techniques based on data mining and natural language processing for analysing the consistency of reviews against specific criteria, which can provide an indication of whether a given review is genuine or malicious spam. The goal of this research is to design the equivalent of a lie detector for ratings and reviews.

Analysis and detection of review spam (aka. opinion spam) is a relatively recent research trend, so the literature is still relatively limited, but some studies show promising results. Studies include Benevenuto *et al.* (2009) [1], Lim *et al.* (2010) [15], Gilbert & Karrie (2010) [7], Wu *et al.* [25, 24], Ott *et al.* (2011) [19], and Duan & Liu (2012) [5].

As with traditional security solutions, it is to be expected that attackers will adapt to robustness solutions implemented in TRSs, thereby resulting in a endless cat-and-mouse game. This phenomenon can already be observed with regard to Google's search engine where the first version of the PageRank algorithm was attacked by link spam, which consists of inserting links to a specific page on open web fora such as discussion groups and wikis. A fix to the link spam problem was to introduce the `no-follow` tag in 2005 which from then on has been automatically added by web server software to every URL inserted in publicly editable web pages. The `no-follow` tag instructs search engines to ignore the link, thereby eliminating the effect of link spam. SEO (Search Engine Optimalization) is to influence search engines to get the highest possible position of a specific web page on the SERP (Search Engine Results Page). In SEO, almost anything goes, and search engines such as Google and Bing must constantly change the way their SERP ranking is computed, in order to prevent strategic manipulation.

2.1 Regulatory and Social Context for Online Reputation

It is important to look beyond purely technical aspects of TRS robustness when prevention of TRS manipulation is the goal. A real TRS is always embedded in a real community or market with its policy and legal context. Even if there are no technical barriers to manipulating a TRS, the fact that it is forbidden by policy or legislation might have a significant preventive effect on potential attackers. As an analogy, there is nothing that physically can stop a car driver for speeding if he really wants to do it. However, the possibility of a fine, or simply knowing the danger that it poses to people

is sufficient to prevent most motorists from speeding. As an alternative to technical robustness mechanisms, it could be useful to define adequate regulation and policies for the deployment and usage of TRSs in online communities.

The simplicity of manipulating TRSs is in many ways paradoxical when considering that that TRSs often have considerable impact on economic performance. A hotel owner can be tempted to use a TRS to gain an advantage over competitors and to maximise profit in the following ways:

(a) Write false positive reviews and artificially inflate own reputation score.
(b) Write false negative reviews and give unfairly negative ratings to competitors.

While activity (a) would appear unethical to most people it would be difficult to define it as directly illegal. Activity (b) on the other hand would not only be considered unethical, but could be considered illegal under most jurisdictions on the basis of legislation regarding defamation.

Defamation is when someone makes a false claim implied to be true which may give a negative image to a person, business, product, group, government, or nation. In order for a complaint against defamation to succeed it is normally required that the defaming claim can be proven false and that the claim is communicated to someone other than the defamed entity. Slander and libel are specific categories of defamation, where slander typically refers to a malicious, false, and defamatory spoken statements, while libel refers to any other form of communication such as written words or images. Online defamation can therefore be characterised as libel. Most jurisdictions allow legal actions, civil and/or criminal, to deter various kinds of defamation and retaliate against groundless criticism.

In the case of Roger M. Grace vs. eBay (2004) [18] the plaintiff, Roger Grace, an eBay buyer, sued eBay and the seller Tim Neely after the seller had posted negative comments about Grace. According to court filings, Neely's comments about Grace were: *"Complaint: SHOULD BE BANNED FROM EBAY!!!! DISHONEST ALL THE WAY!!!!"*. The Superior Court of Los Angeles County held that Section 230 of the US Communications Decency Act of 1996 and the User Agreement on eBay's Web site relieve eBay of liability for libel with respect to comments posted by a seller on the eBay Web site. The user agreement on eBay's website contained the the following section: *"Because we are a venue, in the event that you have a dispute with one or more users, you release eBay (and our officers, directors...) from claims, demands and damages (actual and consequential) of every kind and nature, known and unknown, suspected and unsuspected, disclosed and undisclosed, arising out of or in any way connected with such disputes"*. The court also dismissed the suit against the seller Neely after eBay removed the challenged comments from its website.

While the above case released the owner of the TRS itself from liability, it does leave open the possibility of upholding complaints of libel against the party who produces an alleged defaming statement. Leaving baseless negative feedback and reviews can thus lead to legal prosecution. Not only that, even when users genuinely feel that there is an objective basis for leaving negative reviews, the user still faces the risk of legal action from the target of the negative reviews. This creates risk for anybody who wants to leave negative feedback, which by itself represents a disincentive against leaving negative feedback, even when it is warranted.

TRSs are so widespread in online communities and markets that one can speak about the *reputation society* as a new significant dimension of modern society [17]. Reputation is an asset that can be won and lost, just like real money. We have strict laws governing how money is exchanged, but very little legal regulation regarding reputation. While legislation about defamation provides protection against unfair damage to reputation, there seems to be no typical legislation against unfair inflation of own reputation. From a general point of view, unfair inflation of own reputation can have a negative economic impact on other parties similarly to damaging their reputation. One could therefore argue that there currently is a hole in most legislations in that respect. Participants in online communities thus face little risk when engaging in unfair inflation of own reputation. It is then up to the TRS owner to define specific policies and sanctions against this practice.

Since TRSs often cannot be considered robust, it seems surprising that they still can provide significant value and that they have become so widespread. One might therefore say that TRSs follow the paradoxical "Yhprums Law," which is the inverse of Murphys Law, expressed by: *"Something that shouldnt work sometimes does work."*.

One possible explanation of why TRSs are useful despite their weaknesses is that in many situations, a TRS does not necessarily need to be robust. Resnick & Zeckhauser (2002) [20] consider two explanations: (a) Even though a reputation system is not robust it might serve its purpose of providing an incentive for good behaviour if the participants think it works, and (b) even though the system might not work well in the statistical normative sense, it may function successfully if it reacts swiftly to bad behavior and imposes costs for a participant to get established.

Finally, it could be argued that the TRS in an online community serves as a kind of social glue. A TRS provides an interface through which participants can communicate and relate to each other, which in itself is valuable. Any TRS with user participation will depend on how people can use it to better connect to other participants and to the community as a whole, and must be designed with that perspective in mind.

3 Conclusion

The online world is somewhat analogous to the US Wild West of the 19th century where legislation was unclear and law enforcement was weak. In this context of relative lawlessness, trust and reputation systems represent alternative methods for moderating and regulating online communities. However, the informal and collaborative mechanisms of trust and reputation systems will inevitably come under pressure and attack whenever there is significant financial or political value at stake. In that case, malicious manipulation of a reputation system can only be prevented or mitigated if either 1) there exists regulation or policy that prohibits malicious manipulation with credible sanctioning options, or 2) there are technical mechanisms that can detect and block manipulation attempts. Ideally, both protection principles should be implemented simultaneously. In addition, adequate security mechanisms must be in place in order to prevent hacking attempts against trust and reputation systems or against participants' networks. If adequate robustness can be achieved, well functioning trust and reputation systems will become catalysts for healthy growth in online markets and communities.

References

1. Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 620–627. ACM New York, 2009.
2. A. Clausen. The Cost of Attack of PageRank. In *Proceedings of The International Conference on Agents, Web Technologies and Internet Commerce (IAWTIC'2004)*, Gold Coast, July 2004.
3. C. Dellarocas. Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior. In *ACM Conference on Electronic Commerce*, pages 150–157, 2000.
4. Federico Dini and Giancarlo Spagnolo. Buiyng reputation on eBay: Do recent changes help? *International Journal of Electronic Business*, 7(6):581–598, 2009.
5. Huiying Duan and Feifei Liu. Building and Managing Reputation in the Environment of Chinese E-commerce: A Case Study on Taobao. In Rajendra Kerkar and Costin Badica, editors, *Proc. of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS'12)*, Craiova, Romania, June 2012.
6. Randy Farmer and Bryce Glass. *Building Web Reputation Systems*. O'Reilly Media / Yahoo Press, March 2010.
7. Eric Gilbert and Karrie Karahalios. Understanding deja reviewers. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 225–228. ACM New York, 2010.
8. Jennifer Golbeck. Trust on the World Wide Web: A Survey. *Foundations and Trends in Web Science*, 1(2), 2008.
9. Kevin Hoffman, David Zage, and Christina Nita-Rotaru. A Survey of Attack and Defense Techniques for Reputation Systems (to appear). *ACM Computing Surveys*, 42(1), December 2009.
10. A. Jøsang. Online Reputation Systems for the Health Sector. *electronic Journal of Health Informatics*, 3(1):e8, 2008.
11. A. Jøsang and J. Golbeck. Challenges for Robust of Trust and Reputation Systems. In *Proceedings of the 5th International Workshop on Security and Trust Management (STM 2009)*, Saint Malo, September 2009.
12. A. Jøsang, R. Ismail, and C. Boyd. A Survey of Trust and Reputation Systems for Online Service Provision. *Decision Support Systems*, 43(2):618–644, 2007.
13. Reid Kerr. Smart Cheaters Do Prosper: Defeating Trust and Reputation Systems. In *Proceedings of the 8th Int. Joint Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, July 2009.
14. Reid Kerr and Robin Cohen. An Experimental Testbed for Evaluation of Trust and Reputation Systems. In *Proceedings of the Third IFIP International Conference on Trust Management (IFIPTM'09)*, June 2009.
15. Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady W. Lauw. Detecting Product Review Spammers Using Rating Behaviors. In Jimmy Huang et al., editors, *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948. ACM, New York, Toronto, October 2010.
16. Stephen Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, 1994.
17. Hassan Masum, Craig Newmark, and Mark Tovey. *The Reputation Society: How Online Opinions Are Reshaping the Offline World*. The Information Society Series. MIT Press, 2012.

18. Court of appeal of the state of California. Roger M. Grace vs. eBay. B168765, Los Angeles County, Super. Ct. No. BS288836, 2004.
19. Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319. Association for Computational Linguistics, 2011.
20. 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.* Elsevier Science, 2002.
21. P. Resnick, R. Zeckhauser, R. Friedman, and K. Kuwabara. Reputation Systems. *Communications of the ACM*, 43(12):45–48, December 2000.
22. P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood. The Value of Reputation on eBay: A Controlled Experiment. *Experimental Economics*, 9(2):79–101, 2006. Avaliable from http://www.si.umich.edu/~presnick/papers/postcards/PostcardsFinalPrePub.pdf.
23. A. Withby, A. Jøsang, and J. Indulska. Filtering Out Unfair Ratings in Bayesian Reputation Systems. In *Proceedings of the 7th Int. Workshop on Trust in Agent Societies (at AAMAS'04)*. ACM, 2004.
24. Guangyu Wu, Derek Greene, and Pádraig Cunningham. Merging multiple criteria to identify suspicious reviews. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 241–244. ACM New York, 2010.
25. Guangyu Wu, Derek Greene, Barry Smyth, and Pádraig Cunningham. Distortion as a validation criterion in the identification of suspicious reviews. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 10–13. ACM New York, 2010.
26. B. Yu and M.P. Singh. Detecting Deception in Reputation Management. In *Proceedings of the Second Int. Joint Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, pages 73–80. ACM, 2003.
27. Hui Zhang et al. Making Eigenvector-Based Reputation Systems Robust to Collusion. In *Proceedings of the Third International Workshop on Algorithms and Models for theWeb-Graph (WAW2004)*, Rome, October 2004.