A User-Centric Study Of Reputation Metrics in Online Communities

Stephan Hammer, Rolf Kiefhaber, Matthias Redlin, Elisabeth Andre, and Theo Ungerer

Department of Computer Science, Augsburg University,
Universitaetsstr.6a, 86159 Augsburg, Germany
{hammer,andre}@hcm-lab.de
{kiefhaber,ungerer}@informatik.uni-augsburg.de
redlin@student.uni-augsburg.de

Abstract. With the growing importance of online markets and communities, users increasingly have to interact with unknown people. When choosing their interaction partners, they often lack direct experience and are forced to rely on ratings provided by others who are often unknown themselves. A number of reputation systems have been developed with the aim of improving the credibility of inferred reputation values. Most of these reputation systems proved their accuracy and robustness against manipulation in evaluations and therefore are believed to enhance the users trust in the system. However, what also matters is the users' experience with the reputation system. To investigate whether a reputation systems good functionality is sufficient to enhance the users' rating behavior and the users' trust in the provided reputation values and therefore also the entire system two substantially different reputation metrics were evaluated in an experimental game. The results obtained by this user-centric study are presented in this paper.

Keywords: Trust, Reputation Systems, User Study

1 Introduction

Today users interact in all kinds of online communities. They look for ratings for hotels, products or even experts, such as physicians. They trade in online marketplaces like eBay. They outsource tasks, such as the labeling of data, to online communities[1] and they arrange real-world interactions like carpooling[2] or small jobs like house cleaning or even babysitting[3]. In such communities users mostly have to interact with strangers. Therefore, it is crucial that they can trust in the benevolence and abilities of possible interaction partners. This reduces users' feeling of insecurity and risk [1] and increases their willingness to interact with unknown people [2].

[1] https://www.mturk.com/
[2] http://www.avego.com/
[3] https://www.taskrabbit.com/

The traditional approach of gathering information about someone's reputation entails asking only a small number of trusted people. This results in a small amount of information, but also in mostly credible information. In contrast, today's online approaches include a lot of information provided by a lot of mostly unknown people and thus the users are faced with uncertainty as to whether this information is reliable. Therefore, several reputation metrics, such as [3, 4] were presented to make inferred reputation more credible. All of these reputation metrics were evaluated on their accuracy, e.g., on the Epinions.com database [5], and proved their ability to overcome problems, such as manipulative ratings. Therefore, one could assume that the users trust more in these systems than in simpler ones. However, based on [6] it also matters how users think a reputation metric works and, more importantly, that users trust in the entire system's reliability, even if they do not know how it works. Therefore, two versions of an experimental game with substantially different reputation metrics, the Neighbor-Trust Metric (NTM) [7] and eBay's reputation metric[4], were designed. These versions were utilized in a user-centric study to investigate whether a reputation system's good functionality is sufficient to enhance users' rating behavior and users' trust in the reputation values provided and therefore the entire system. This paper presents the results gained from this study and possible steps to improve the users' experience with reputation systems.

The remainder of the paper is structured as follows: Section 2 gives a short evaluation of different reputation metrics from users' point of view. In Section 3 we introduce the most important aspects of the reputation metrics, eBay's reputation metric and the NTM, that were compared in the user study. The experimental game, the user study conducted with the game and a discussion of the results and experiences are presented in Section 4. Section 5 concludes the paper and presents future work.

2 Reputation Metrics

Because trust between interacting and cooperating subjects is a major issue in many fields of research several reputation metrics already exist. In general they are divided into global and local metrics. In this section, they are compared from a user's point of view.

Global reputation metrics, such as eBay's reputation metric, infer a unique global reputation for every user and do not take into account subjective perceptions of users. This is contrary to the diverse characters and opinions of all kinds of people that take part in online communities. If inferring the reputation of users with a lot of ratings this seems to be no problem, because the global reputation consists of many ratings provided by diverse users and therefore generally fits most of the users' opinion. Furthermore, users that received many ratings, in general, also received mostly positive ratings. However, for users that received only a small number of ratings it is difficult to infer, if the assessed reputation

[4] http://pages.ebay.com/help/feedback/scores-reputation.html

will match the actual experience. Since most of the users in online communities only received few ratings [6], this is a big issue.

In comparison to that, local metrics take into account that users' opinions on others' statements or trustworthiness can differ and are very subjective. To assess the trustworthiness of so far unknown users, TidalTrust [3] and Moletrust [5], for example, take into account that people feel more confident about information provided by known and trustworthy people than about information provided by unknown people. Therefore, they include only ratings provided by trustworthy users. That again is a problem, if we think about the reality in online communities in which users often have to interact with people that probably are unknown to the users' trusted people, too. In this case, a user assesses people's benevolence, competence or trustworthiness without any provided information.

Other metrics, such as the FIRE metric [11], consider the ratings provided by all former interaction partners of the target user. However, without a mechanism that verifies the accuracy of these trust statements, this approach is vulnerable to attacks and manipulations. Malicious participants or groups, for example, could offer false ratings to promote untrustworthy partners or blur the reputation of other users [10].

The Eigen-Trust metric [4] as well as the Neighbor-Trust Metric (NTM) [7] enhance this approach by the identification and isolation of manipulating participants. Thereby, both are able to infer the reputation for unknown participants based on the assessment of trusted as well as of unknown participants in a trustworthy way. However, the NTM extends the Eigen-Trust metric by separating the trust values for the direct interaction between users and for the reputation users provide about each other. The reason for this is, that a bad interaction partner nevertheless could be a good informant and vice versa. The details of this approach will be explained in Section 3.

3 The Evaluated Reputation Metrics

For the study, two substantially different reputation metrics were chosen to investigate whether a reputation system's good functionality is sufficient to enhance users' rating behavior and users' trust in the reputation values provided and therefore the entire system.

Since eBay's Marketplaces ended the first quarter 2013 with 116 million active users[5], eBay's reputation metric[6], despite the already mentioned drawbacks, is one of the best-known reputation metrics and seems to be accepted by the users. Furthermore, it is also one of the few metrics to be analyzed with regard to their influence on the users' behavior [8, 9]. The results of these studies, for example, showed that only half of all trades on eBay were rated and that the majority of provided ratings were positive. Although at first sight the last result could be interpreted as a success, a closer look at the data revealed two problems: because there was a high correlation between the ratings provided by buyers and sellers,

[5] http://investor.ebay.com/releasedetail.cfm?ReleaseID=757272
[6] http://pages.ebay.com/help/feedback/scores-reputation.html

Resnick and Zeckhauser supposed that the users (1) reciprocated and (2) feared retaliation [6]. To address these problems, sellers no longer are allowed to give negative or even neutral ratings, so as to alleviate buyers' fears of retaliation or unfair ratings. Instead, sellers can only leave comments on unfair ratings and can request a revision of the rating by the buyer[7]. This does not seem to be a trustworthy approach to handle possible manipulations of ratings. However, the users seem to accept the reputation system. Therefore, it was chosen to be one of the utilized reputation metrics in the study.

In comparison, our Neighbor-Trust Metric (NTM) [7] gathers the direct trust values t_{ic} from all former interaction partners i of a target user c, called "neighbors", and aggregates them by a weighted mean metric to assess an individual, local reputation value r_{ac} for every user a:

$$r_{ac} = \frac{\sum_{i \in neighbors(c)} w_{ai} \cdot t_{ic}}{\sum_{i \in neighbors(c)} w_{ai}}$$

The weights w_{ai} represent the trust of the user a in the trust values the neighbors i provide. The reason for the separation of the trust values for direct interactions between users and for trust ratings users provide to each other is, that a bad interaction partner could nevertheless be a good informant, and vice versa. The weights are adapted after every interaction. When a user a had a direct experience with a user c and provided a trust rating t_{ac}, this rating is compared to the trust rating t_{bc} a user b provided before the interaction. If b gave information that corresponded with a's own experience, then the future statements of b will be weighted higher than before. Correspondingly, if the ratings differ, the weight will be lowered. Thus, the metric is not only able to learn about the trustworthiness of the interaction partner, but also to identify users that provide false or non-conformist ratings. Furthermore, by weighting down these users' ratings, inferred reputation values later will be more trustworthy and accurate. Therefore, by overcoming the vulnerability to manipulation the NTM should be more trustworthy for users than, for example, eBay's metric.

4 The User Study

4.1 Experimental Design

We investigated the influences of different reputation metrics on users' trust and rating behavior by comparing two versions of an experimental game that was inspired by other experimental games [12–14]. The two versions of the game differed only in the utilized reputation metrics. One version used eBay's metric and the other version used the NTM. We believed that two results could be possible: (1) The NTM's robustness against manipulation and the more credible reputation values (A) increase users' trust in the system and (B) cause more

[7] http://pages.ebay.com/help/feedback/feedback-disputes.html

honest ratings. (2) There is no difference in using eBay's metric or the NTM, for instance, because users do not recognize the different functionalities, since there are too short and too few interactions in the game and in online communities in general.

The experimental game was designed as a collaborative quiz (see Fig. 1). We assumed that collecting points and the chance to win prizes would be engaging and emotive.

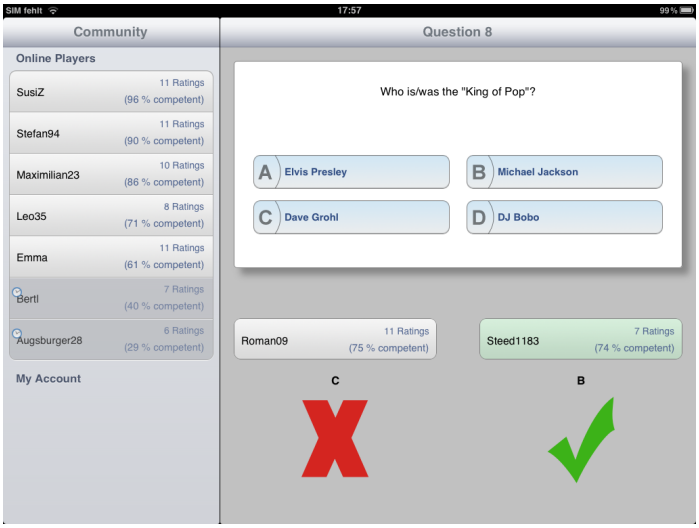

Fig. 1. Collaborative quiz

To enable a realistic comparison with cooperations in online communities, the following process sequence was designed:

1. A user has to choose an interaction partner. (In the study the interaction partners (teammates), were simulated by seven virtual players (VP) that were available from the beginning and had reputation values (RV) from 40% to 100%.)
2. The requested user has to confirm the cooperation. (In the study the decision of the VPs depended on their own reputation and the reputation of the requesting user. The user was rejected if her reputation was 20% lower than the VPs reputation (see Fig. 2 top)).
3. An interaction is successful if both users complete it successfully. (That is, both players have to answer a question correctly to get a point. The probability of a correct answer by a VP was RV - 10% for easy questions and RV - 30% for difficult questions. Therefore, players with a high reputation answered correctly more often than players with a low reputation.)

4. To increase all users' chances of gaining a higher benefit, interaction partners have to rate each other after each cooperation. This enables all users to distinguish between good and bad interaction partners.
5. When starting a new interaction, each user has to choose an interaction partner again. (Since the users were allowed to choose the same VP again, a VP that was chosen three times in a row entered an "idle" state, to prevent the participants from choosing the same VP throughout the entire study (see Fig. 3). This status lasted for three rounds.)

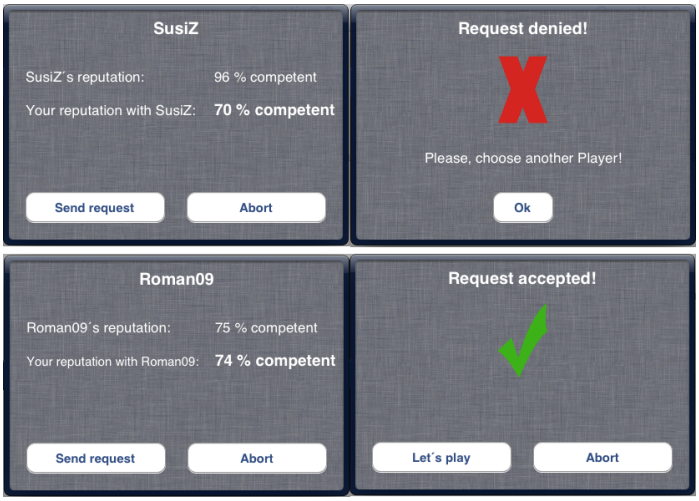

Fig. 2. Confirmation of user request depending on reputation. Top: rejection, bottom: acceptance

To investigate the users' reactions in different situations, based on [6], a variety of hardcoded behaviors for the VPs was implemented: (1) In general, the rating of the VPs corresponded to the user's answer. (2) If a user answered wrongly and rated the VP positively (independently of his answer), some VPs returned this favor and rated positively, too. (3) In a few cases some VPs rated a user negatively out of revenge if they received a negative rating.

4.2 Experimental Setting

Both versions of the quiz were played by half of the participants. In both versions the participants had to answer the same 10 easy and 10 difficult general knowledge questions. Based on the results of [6], such a small number of interactions corresponds to the actual conditions in online communities like eBay. Furthermore, it can be assumed that users that do not establish trust in a system during the first interactions will not use the system. The total number of

received ratings in both versions of the quiz was shown for every player (see Fig. 3). Additionally, in the eBay-version a global unique reputation value, equal to that provided on eBay, was shown to support the user's selection of the next teammate (see Fig. 3 left). In comparison, in the NTM-version an individual local reputation value calculated by the NTM was shown (see Fig. 3 right).

To analyze the accuracy of the provided ratings and the users' selection of their teammates, the names of the chosen teammates, and the answers and ratings of the user and the current teammate for each question were logged. Moreover, interesting behavior was documented by hand. To analyze the participants' experiences with the respective quiz-version, they had to fill in questionnaires after they completed the quiz.

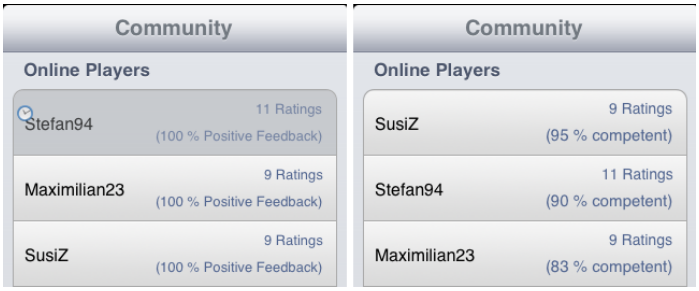

Fig. 3. Ranking of Virtual Players. left: eBay's reputation metric; right: Neighbor-Trust Metric

4.3 Conducting the Study

At first, the users had to fill in a questionnaire to provide general demographic information, and information about experiences with strangers and rating systems on the internet. Then, after a short introduction, the users had to play the quiz. The users were not informed about the functionality of the rating system. To increase the participants' ambition, they were told that they could win prizes depending on their results. After the quiz, the participants had to rate statements concerning their experiences with their teammates and the utilized rating system. All statements in the questionnaires had to be rated on a Likert scale from 1 ("not at all") to 5 ("definitely"). Ratings lower than 3 were interpreted as disagreement with a statement and ratings higher than 3 were interpreted as agreement.

4.4 Results and Experiences from the Experimental Game

Overall 16 women and 26 men aged between 22 and 56 (mean: 31.5) took part in the user study. The participants studied and worked in all kind of professions related (43%) and not related (57%) to computer science.

All participants already had interacted with unknown persons, e.g. on eBay, or had trusted in reviews on products or holiday destinations. Asked for their frequency of interactions with unknown people, the largest proportion of participants answered with "several times a year (29%)" or "several times a month (26%)". More than half of the participants reported on good (45%) or excellent (12%) experiences with unknown persons. All other participants rated their experiences as "neutral" and explained their ratings, for example, with mediocre information provided by others. Most of the participants agreed with the statement "Whenever you meet strangers, you have to be on guard until they have proven that they are trustworthy.". The average rating (M) was 3.62 (Standard Deviation (SD) = 0.90). This matches the fact that most of them also considered rating systems important (M=3.67; SD=0.89), because they allow an objective assessment of the trustworthiness of unknown people and decrease the chances of negative experiences. However, half of the participants were in doubt about the honesty of the provided feedbacks and some criticized possible manipulation and the lack of transparency of reputation systems. Nevertheless, several of the participants declared that reputation systems at least provide an indication of a user's trustworthiness.

A two-sided dependent t-test showed no significant differences for users' trust in the utilized reputation systems. Neither users' trust towards the provided ratings (NTM: M = 3.19; SD = 0.73; eBay: M = 2.95; SD = 0.72; p = 0.31) nor the perceived usefulness of the reputations systems (NTM: M = 3.80; SD = 0.66; eBay: M = 3.85; SD = 0.55; p=0.81) suggest that the NTM's ability to identify false ratings was recognized by the users. This was confirmed by an average rating of 1.71 (SD=0.76) when asked if they believed that the system is able to identify false ratings (eBay: M=2.05; SD=1.05). However, in both versions of the quiz almost all users stated that they based their selection of the teammates on the provided reputation values. But 67% of all participants also showed confidence and repeatedly selected players with whom they already had positive experiences, such as right answers or generous ratings and half of the users even based their choices mainly on positive experiences. This matches the results in [12] that direct and repeated interactions between users are the primary reason for increased trust.

The comparison of users' rating behavior in the two versions of the quiz showed small differences (see Fig. 4). But since the users did not recognize the NTM's ability to identify false ratings, these small differences seem not to be caused by the utilized reputation metrics. In both versions the participants rated honestly and rated positively if their teammates gave a correct answer (NTM: in 98% of the cases; eBays: 97%) and rated negatively if their teammates gave a wrong answer to easy questions (NTM: 73%; eBays: 65%) (see Fig. 4 (top)). However, there were many generous ratings (NTM: 50%; eBays: 55%) if the VPs gave wrong answers to difficult questions (see Fig. 4 (bottom)). Some of the users explained overly good ratings in general by the saying "To err is human". Furthermore, 33% of the participants admitted that they reciprocated, because of former positive experiences with the regarding teammate, such as

right answers or prior generous ratings towards themselves. In this regard, 79% of all participants agreed that users in online communities can be convinced to rate positively if they received a positive rating in return (M=3.83; SD=0.65). 20% of all participants also explained that they provided overly good ratings because they feared retaliation. However, half of the participants negated that they would fear retaliation in general and the average score was 2.95 (SD=1.0). In summary, almost half of all wrong answers by the VPs were rated neutral or even positively.

Ratings (NTM-Users)
Easy Questions:

	positive	neutral	negative
right	146	2	0
false	4	13	45

Difficult Questions:

	positive	neutral	negative
right	100	2	0
false	9	45	54

Ratings (eBay-Users)
Easy Questions:

	positive	neutral	negative
right	151	2	0
false	4	16	37

Difficult Questions:

	positive	neutral	negative
right	106	4	2
false	10	43	45

Fig. 4. Ratings provided for right and false answers. NTM-Users (left) and eBays-Users (right); Easy Questions (top) and Difficult Questions (bottom)

5 Conclusion

This paper presented an experimental game by which two substantially different reputation metrics, Neighbor-Trust Metric (NTM) and eBay's reputation metric, were investigated from a user-centric perspective. The comparison of the metrics showed only small differences for users' rating behavior and no significant differences concerning users' trust in the reputation systems. This indicates that accuracy and robustness against manipulation are not the only criterions for good reputation systems. In addition to the general vulnerability of rating systems to manipulation, most of the participants in the study criticized the lack of transparency of rating systems. An improved transparency could therefore enhance users' experience of reputation systems. For reputation systems, such as the NTM, which assesses the credibility of a user's rating behavior, it could be a good idea to display this additional information. It could help to explain the inferred reputation and thus users' trust in the assessed reputation values could be increased. Furthermore, an additional criterion would be introduced that could support the choice of future interaction partners based on users' preference to interact with people that provide honest ratings. Finally, the amount of overly good ratings could be reduced, too.

Acknowledgments This research is sponsored by *OC-Trust* (FOR 1085) of the German research foundation (DFG).

References

1. McKnight, D.H., Choudhury, V., Kacmar, C.: Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. Info. Sys. Research. 13(3), 334–359 (2002)
2. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An Integrative Model of Organizational Trust. The Academy of Management Review. 20(3), 709-734 (1995)
3. Golbeck, J.: Computing and Applying Trust in Web-Based Social Networks. Ph.D. Dissertation, University of Maryland at College Park, College Park (2005)
4. Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H.: The Eigentrust Algorithm for Reputation Management in P2P Networks. In: 12th international conference on World Wide Web, pp. 640–651. ACM, New York (2003)
5. Massa, P., Avesani, P.: Trust-aware Recommender Systems. In: 2007 ACM conference on Recommender systems, pp. 17–24. ACM, New York (2007)
6. Resnick, P., Zeckhauser, R.: Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System. In: Baye, M.R. (ed.) The Economics of the Internet and E-Commerce 2002. vol. 11, pp. 127-157. Elsevier Science, Amsterdam (2002)
7. Kiefhaber, R., Hammer, S., Savs, B., Schmitt, J., Roth, M., Kluge, F., Andre, E., Ungerer, T.: The Neighbor-Trust Metric to Measure Reputation in Organic Computing Systems. In: 5th IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshops, pp. 41–46. IEEE Press, New York (2011)
8. Resnick, P., Zeckhauser, R., Swanson, J., Lockwood, K.: The Value of Reputation on eBay: A Controlled Experiment. Experimental Economics. 9(2), 79-101 (2006)
9. Canals-Cerd, J.: The Value of a Good Reputation Online: An Application To Art Auctions. Journal of Cultural Economics. 36(1), 67–85 (2012)
10. Dellarocas, C.: Immunizing Online Reputation Reporting Systems Against Unfair Ratings And Discriminatory Behavior. In: 2nd ACM conference on Electronic commerce, pp. 150–157. ACM, New York (2000)
11. Huynh, T.D., Jennings, N.R., Shadbolt, N.R.: An Integrated Trust and Reputation Model for Open Multi-Agent Systems. In: Autonomous Agents and Multi-Agent Systems. 13(2), 119–154 (2006)
12. Bolton, G.E., Katok, E., Ockenfels, A.: How Effective Are Electronic Reputation Mechanisms? An Experimental Investigation. Management Science. 50(11), 1587–1602 (2004)
13. Bohnet, I., Harmgart, H., Huck, S., Tyran, J.-R.: Learning Trust. Journal of the European Economic Association. 3(2-3), 322–329 (2005)
14. Keser, C.: Experimental Games for the Design of Reputation Management Systems. IBM Syst. J. 42(3), 498–506 (2003)