VISUALIZING UNFAIR RATINGS IN ONLINE REPUTATION SYSTEMS

Complete Research

Sänger, Johannes, University of Regensburg, Regensburg, Germany, johannes.saenger@ur.de
Richthammer, Christian, University of Regensburg, Regensburg, Germany, christian.richthammer@ur.de
Kunz, Michael, University of Regensburg, Regensburg, Germany, michael.kunz@ur.de
Meier, Stefan, University of Regensburg, Regensburg, Germany, stefan.meier@ur.de
Pernul, Günther, University of Regensburg, Regensburg, Germany, guenther.pernul@ur.de

Abstract

Reputation systems provide a valuable method to measure the trustworthiness of sellers or the quality of products in an e-commerce environment. Due to their economic importance, reputation systems are subject to many attacks. A common problem are unfair ratings which are used to unfairly increase or decrease the reputation of an entity. Although being of high practical relevance, unfair rating attacks have only rarely been considered in literature. The few approaches that have been proposed are furthermore quite non-transparent to the user. In this work, we employ visual analytics to identify colluding digital identities. The ultimate benefit of our approach is the transparent revelation of the true reputation of an entity by interactively using both endogenous and exogenous discounting methods. We thereto introduce a generic conceptual design of a visual analytics component that is independent of the underlying reputation system. We then describe how this concept was implemented in a software prototype. Subsequently, we demonstrate its proper functioning by means of an empirical study based on two real-world datasets from eBay and Epinions. Overall, we show that our approach notably enhances transparency, bares an enormous potential and might thus lead to substantially more robust reputation systems and enhanced user experience.

Keywords: Trust, reputation system, unfair ratings, collusion, visual analytics.

1 Introduction

Modern e-commerce platforms such as electronic marketplaces provide a valuable environment that brings together millions of actors to trade goods and services. Buyers and sellers are thereby offered unprecedented opportunities involving an almost infinite variety of products. Regardless of whether a buyer is looking for antiquarian books, brand-new technologies or highly specialized instruments, he will find a suitable transaction partner on the Web in most of the times. However, this "universe of strangers" also poses many challenges (Dellarocas, 2006). Unlike in traditional face-to-face transactions, buyers do neither get a complete picture of the products' actual quality nor do they know about the trustworthiness of a seller. Since advance payment is a usual practice in multiple settings, buyers often face high risks. To cope with this challenge, many e-commerce systems encourage customers to provide feedback after a transaction denoting their satisfaction. Reputation systems collect all evidence, aggregate the input data and provide one or several reputation values as output. In this way, reputation systems can support buyers in deciding whom to trust and what product or service to choose.

According to a recent study carried out by Diekmann et al. (2014), sellers with better reputation obtain higher prices and have an increased number of sales. While on the one hand promoting trustworthy participation, this also bares an incentive for malicious actors to unfairly push their reputation in order to gain more profit. Unfair ratings have already become a serious problem in practice. Favorably recommending each other to dishonestly increase the reputation on eBay[1] or slandering on review sites to unfairly destroy the reputation of a competitor such as on TripAdvisor[2] are only a few of multiple examples (Jøsang and Ismail, 2002; TripeAdvisor, 2012).

The problem of unfair ratings has therefore been a topical issue in research for many years. Solutions introduced to deal with this problem span statistical filtering, clustering-based discounting and evaluation techniques using hidden markov chains (Sänger and Pernul, 2014a). These approaches, however, are highly non-transparent to end users since a numerical value as outcome cannot convey any detailed information about the input data and the calculation techniques used. To address this challenge, we make use of visual analytics. Visual analytics is based on the principle to incorporate the cognitive capabilities of a human analyst and the computing power of a machine. By means of a software tool[3] we implemented, we will demonstrate that using interactive visualizations, collusive unfair rating attacks on reputation systems can be reliably and transparently detected and filtered. We will furthermore show that our approach allows to combine endogenous and exogenous detecting mechanisms to mitigate their respective weaknesses.

The remainder of this paper is organized according to the guidelines for conducting design science research proposed by Hevner et al. (2004). We thereby follow the "nominal process model" introduced by Peffers et al. (2007). Firstly, we describe the background and related work that is important with respect to our approach in Section 2. Here, we motivate our work, delineate the research gap we discovered and define the objectives of our solution ("Problem identification and motivation" and "Objectives of a solution"). Secondly, we propose our interactive visualization-based approach in Section 3. We thereby describe the conceptual design as a generic mechanism that can be adapted to a specific application area. Then, we show how the concept was implemented in a prototype ("Design and development"). Subsequently, we demonstrate the proper functioning of our prototype using real-world datasets from eBay and Epinions[4] in Section 4. Here, we carry out two case studies in which we show how to reliably identity unfair rating attacks using our prototype and evaluate the outcomes ("Demonstration" and "Evaluation"). Finally, we sum up our findings and conclude in Section 5.

[1] http://www.ebay.com/
[2] http://www.tripadvisor.com/
[3] http://trust.bayforsec.de/unfair_ratings/
[4] http://www.epinions.com/

2 Background

In this section, we give an overview of the research areas that are relevant with respect to our work. Firstly, we introduce the fundamental concepts of online trust and reputation systems. Secondly, we outline related work regarding unfair rating attacks and defense mechanisms. We thereby point out the weaknesses of common solution approaches and motivate our work. At the same time, the concept of visual analytics will be introduced. Based on this, we clarify the research gap and define the solution objectives.

2.1 Online Trust and Reputation Systems

The notion of trust has been discussed in research for decades. It is not only important for the computer science community but also for various other research domains such as sociology, psychology, economics, philosophy, and media science. Consequently, there are just as many interpretations of it and a universally accepted definition is still missing. As our work particularly focuses on online trust, we refer to Gambetta's (1988) definition of *reliability trust* that is commonly cited in the literature on online trust and reputation. In short, the definition regards trust as the subjective probability with which the entity under observation assesses that another entity will perform a particular action. One way to come up with this probability is through reputation-based (as opposed to policy-based) trust establishment (Artz and Gil, 2007). Reputation-based trust is derived from past interaction experiences and behavior associated with an entity, which includes both one's own perceptions and recommendations by others (Habib et al., 2013).
Since the number of entities involved in an online environment may be of the order of millions, manually determining their reputation becomes unmanageable. To address this, reputation systems have emerged in the recent decades. Reputation systems encourage actors of a community to leave feedback about the behavior of an entity. They then collect all evidence available, aggregate the data and provide one or several reputation values as output. The application areas span multiple fields such as e-commerce (Resnick and Zeckhauser, 2002), P2P networks (Gupta et al., 2003), and virtual organizations (Winkler et al., 2007), just to name a few. In electronic marketplaces, for instance, buyers are able to rate sellers after each transaction. Based on these experiences, future customers can decide whether to trust a seller and as a consequence whether to buy. Hence, high reputation is not only an evidence for trustworthiness but also leads to an increased number of sales and higher prices, as pointed out by Diekmann et al. (2014). This bares an incentive for actors to unfairly push their reputation.

2.2 Unfair Rating Attacks

Because of their economic importance, reputation systems have been subject to various kinds of attacks. So far, most research activities have focused on seller attacks, meaning what an adversary is able to do in the role of the seller. Typical examples for seller attacks include playbooks, value imbalance exploitation, re-entry, discrimination, and reputation lag exploitation, to name the most important. Advisor attacks, in contrast, have received less attention. According to Jøsang and Golbeck (2009), advisor attacks can be summarized under the term "unfair rating attacks" because they are based on one or several digital identities providing unfair ratings to other digital identities. In this work, we distinguish different kinds of advisor attacks according to two dimensions. Firstly, ratings can either be unfairly high or unfairly low. These two types of attacks are also referred to as "ballot stuffing" and "badmouthing", respectively (Dellarocas, 2000). Secondly, advisor attacks can be carried out either by one single digital identity ("one-man show"), by several digital identities focusing on one target ("Sybil attack"), or by several digital identities positively influencing one another ("collusion attack"). Note that the term "Sybil attack" normally refers to one entity creating and maliciously using multiple digital identities. Since we cannot distinguish multiple entities from multiple identities, we use the term in a different way in this work. Regarding the two attack dimensions, it has to be pointed out that not all possible combinations are equally relevant in real-world scenarios. Figure 1 provides an overview on this.

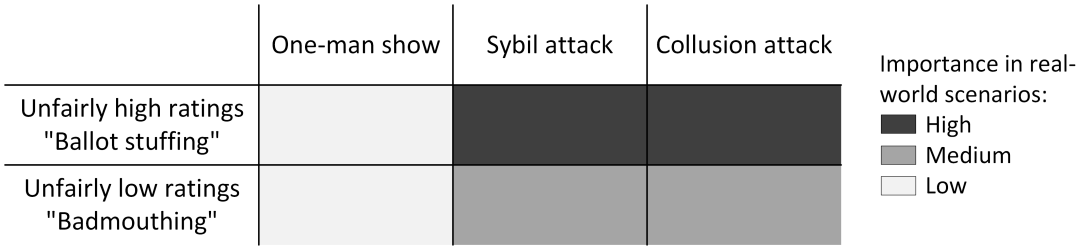

Figure 1. Dimensions of unfair rating attacks and their importance in real-world scenarios.

On the one hand, badmouthing is generally harder to perform than ballot stuffing as most reputation systems used in e-commerce only allow to provide feedback after a successful transaction. Since transactions are bound to costs, there usually is an investment barrier. To push the reputation of a colluding digital identity through unfairly high ratings, in contrast, fake transactions can be used and require a considerably lower investment. Moreover, unfairly low ratings can easily be reported by the actor harmed whereas in ballot stuffing attacks, there is nobody to claim against. On the other hand, attacks performed by multiple digital identities are more important than attacks implemented by single individuals. In the one-man show scenario, badmouthing will quickly lead to other sellers reporting the attacking digital identity. Ballot stuffing, again, does not really make sense because there is no reason for pushing the reputation of a non-colluding seller. If at all, one-man show attacks can only serve as an attack against the system as a whole (e.g. always providing totally random ratings). Taking this into consideration, we mainly focus on unfairly high ratings that are provided through Sybil or collusion attacks.

2.3 Detection of Unfair Rating Attacks

As the problem of unfair ratings was one of the first weaknesses of reputation systems discussed in literature, a few defense mechanisms have been proposed in the recent years. According to Whitby et al. (2004), the body of related work on the detection of unfair rating attacks can broadly be divided into two groups: endogenous discounting and exogenous discounting. Endogenous discounting methods try to detect unfair ratings on the basis of their statistical properties. An early probabilistic approach in this domain was suggested by Dellarocas (2000), who used cluster filtering methods to separate unfair ratings from fair ratings. Whitby et al. (2004) furthermore developed a statistical method based on the beta reputation system introduced by Jøsang and Ismail (2002). They noted that their filter was less effective if more than 30% of the users of the platform were unfair raters, and even counterproductive if the share exceeded 40%. The presence of too many unfair ratings is one of the common problems of majority rule-based approaches. Other early probabilistic approaches were compared by Zhang et al. (2008). A related class of approaches that has come into focus more recently is based on signal modeling (e.g. (Sun and Liu, 2012; Yang et al., 2009)).

Exogenous discounting methods, as the second group of defense mechanisms, are not based on ratings and their statistical properties but focus on the reputation of the raters instead. The underlying assumption is that raters with a bad reputation are more likely to provide dishonest ratings. Cornelli et al. (2002) introduced a reputation scheme on top of the Gnutella P2P network. Users could request opinions about the foreign entities offering a particular file from their peers and, based upon them, decide from which entity to download. Yu and Singh (2003) examined trust networks, especially in e-commerce, and proposed a variant of the weighted majority algorithm (Littlestone and Warmuth, 1994). In their approach, advisors were assigned with weights which could be decreased after unsuccessful predictions. A related strategy often employed in literature is to use ratings provided by a priori trusted agents as a benchmark (Jøsang and Golbeck, 2009). However, these trusted agents have to be determined first. Moreover, the strategy is not feasible under certain circumstances such as in the case of a discriminating seller.

Overall, we found that all known defense mechanisms apply a purely mathematical approach that leads to one or several numerical reputation values as output. These computation methods, however, are non-transparent and incomprehensible for the end user (Hammer et al., 2013). A number as the only output does indeed not reveal many details about the content of the input data. As a consequence, many users become skeptical. For this reason, we pursue a different path by providing an interactive visualization for reputation assessment. By involving the user in the evaluation process, we even enable to interactively switch between endogenous and exogenous techniques to detect unfair ratings.

2.4 Visual Analytics

Visualization is one of the key factors in this proposal. Therefore, we adopt some ideas from visual analytics, which is an interdisciplinary and fast-growing research field combining automated analysis techniques with interactive visualizations. Transferred to our problem scenario, we can employ the computing power of machines to perform clustering operations and the visual-cognitive capabilities of humans to analyze the outcomes of the computations visualized in a meaningful way. We use the visual analytics process introduced by Keim et al. (2010) as the basis for developing our detection methodology (see Section 3.1) as well as the corresponding prototype (see Section 3.2).
Visual analytics has so far not played a role in reputation systems except one previous work addressing the visual detection of seller attacks (Sänger and Pernul, 2014b). Further approaches that tried to include visualizations only made use of static charts or graphs such as the trust network depicted by O'Donovan et al. (2007). Another proposal closely related to this work was presented by Wang and Lu (2006). Although their application scenario was the detection of Sybil attacks in wireless networks and thus different from ours, they had similar goals and ideas. They wanted to design a user-friendly visualization to reveal meaningful correlations and ultimately identify the Sybil attack. Moreover, they argued that the information on the network topology have to be arranged appropriately in order to demonstrate significant features of intrusions. In contrast to our approach, Wang and Lu (2006) assumed that the links among the nodes in the network are bidirectional and thus not directed. They furthermore do not allow multiple edges and edge weights, which makes our visualization technique strongly different.

2.5 Research Gap and Objectives

As pointed out before, most metrics and defense mechanisms to detect unfair ratings used in common reputation systems are quite non-transparent since they only provide an aggregated reputation value that does not reveal any details of how it was determined. In a user-centric study carried out by Hammer et al. (2013), more than half of the participants criticized this lack of transparency. Enhanced transparency could therefore notably increase the user experience in reputation systems. Visualizations are perfectly suitable for depicting a load of data in one picture. Furthermore, they enable to easily and transparently detect structures and anomalies. In this work, we are taking a first step toward detecting unfair ratings by means of interactive visualizations and particularly focus on collusion and Sybil attacks.
In our approach, we want to combine the two broad classes of unfair rating detection methods introduced before. Depending on the attack scenario, either endogenous or exogenous ones can be the more promising choice. Thus, we offer an endogenous discounting component in the form of a clustering algorithm that takes into account the feedback received by digital identities related to the identity in focus. The exogenous discounting component is realized by implementing a sorter that operates on the reputation values of the advisors of the identity in focus. We then let the end user select the appropriate detection techniques.
In a nutshell, our overall goal is to use visual analytics concepts to identify colluding digital identities and enable the end user to filter their unfairly high ratings. The ultimate benefit is the transparent revelation of the true reputation of a particular seller by interactively using both endogenous and exogenous filtering methods.

3 Detecting and Filtering Unfair Ratings through Interactive Visualizations

In this section, we present our novel visualization-based approach to interactively identify and filter unfair ratings in online trust and reputation systems with focus on e-commerce. Firstly, we expose the conceptual design of our visual analytics methodology, which is independent of the reputation model in use. Thus, this concept serves as a generic mechanism that can be adapted to a specific application area. Secondly, we show how the conceptual design was implemented in a software prototype.

3.1 Conceptual Design

Adapting the visual analytics process, the conceptual design of a visual analytics software can be described within the two essential building blocks "models" and "visualization & interaction techniques". The selection of appropriate models and visualization & interaction techniques, however, cannot be made *ex ante* but depends on the structure of the raw data that serve as input. Therefore, before describing the models, an analysis of the input data needs to be carried out in a preliminary step.

3.1.1 Preliminary Analysis

A typical e-commerce setting involves three groups of entities, namely buyers, sellers and products. In electronic marketplaces, actors can be both buyers and sellers. Concerning the trust relationships between those entities, we typically have two roles: the trusting entity (trustor) and the trusted entity (trustee). The buyer usually takes the role of the trustor who wants to establish a trust relationship toward a seller or, if the seller is trusted, toward the quality of a product or service. To support buyers in building such trust relations, other buyers are encouraged to leave feedback denoting their satisfaction. This feedback mostly consists of a numerical rating value and a textual review. Feedback is therefore created as a directed opinion from a buyer toward a seller/product. As this feedback can be either honest or dishonest, many systems furthermore allow to rate the quality of feedback, and thus, the trustworthiness of a buyer as a recommender. In this way, the most reliable feedback/recommender can be determined.
Obviously, there is a difference between feedback given to sellers/products (subsequently referred to as seller feedback) and feedback given to buyers/reviews (subsequently referred to as recommender feedback). Seller feedback might help other actors to decide whether a seller is trustworthy or whether a product is of high quality. Recommender feedback, in contrast, gives a lead whether a recommender is trustworthy in the sense of how reliable and honest his feedback is. Therefore, we distinguish between two types of feedback, seller and recommender feedback. Referrals of each type form a feedback network, the seller feedback network and the recommender feedback network.

3.1.2 Building Block: Models

To model this setting, the feedback networks are represented by two graphs. Throughout this paper, we use the two terms network and graph interchangeably. The seller feedback graph G_s contains all feedback given by buyers toward sellers or products while the recommender feedback graph G_r contains all feedback given toward recommenders. As a pair of nodes can be involved in more than one transaction and as all edges are associated with a numerical rating value, the feedback graphs are defined as weighted directed multigraphs: $G_{s,r} = (V, E, R, f, w)$ where V is a set of vertices representing the entities and E is a set of edges representing the feedback relations. R is the rating set that is formed by all rating values allowed by the reputation system. Furthermore, we have a function $f : E \to V \times V = \{(u,v) | u,v \in V\}$ describing the direction of a feedback relation and a function $w : E \to R$ denoting the rating values associated with each edge.
Figure 2 depicts an example scenario for an electronic marketplace ($R = \{-1, 0, 1\}$). Here, five parties are involved: A who acts as a seller only, B who takes the role of both buyer and seller, and C who is a buyer

only. We furthermore have two nodes D and E who negatively rated the recommendation made by B and positively rated all recommendations made by C.

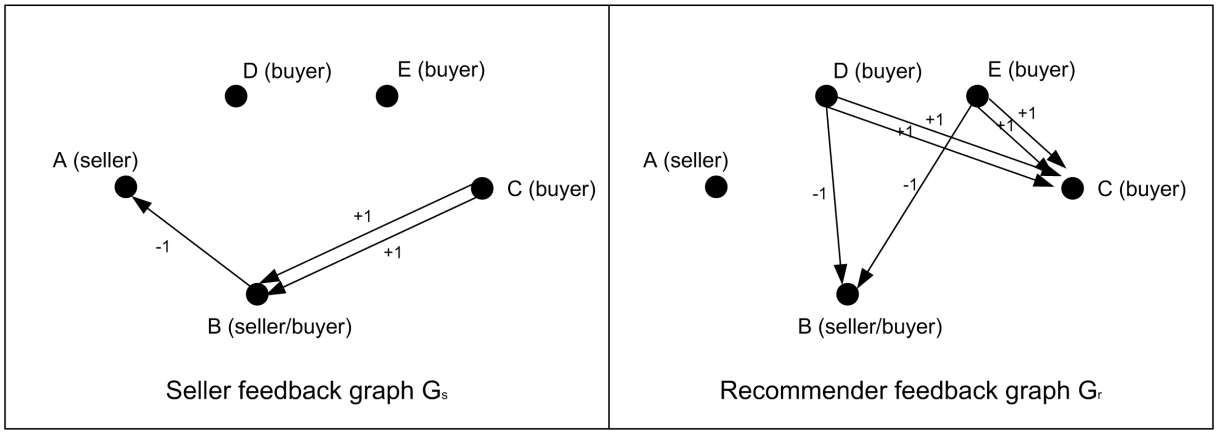

Figure 2. Feedback graphs after three transactions between actors A,B and C.

To analyze the role and position of single nodes in these networks, we set up a list of common graph measures/operations shown in Table 1.

Measure/Operation	**Description**
$VertexInDegree(G,x)$	The number of incoming edges of a vertex x in a network G. The *VertexInDegree* is equivalent to the number of feedback received. Example (for the scenario above): $VertexInDegree(G_s,B) = 2$
$VertexInDegree(G,x,y)$	The number of edges from vertex y to vertex x in a network G. Example: $VertexInDegree(G_r,A,B) = 0$
$AverageVertexInWeight(G,x)$	The average of the weights of all incoming edges of vertex x in the network G. *AverageVertexInWeight* is equivalent to the simple average as a reputation measure. Example: $AverageVertexInWeight(G_r,B) = -1$
$AverageVertexInWeight(G,x,y)$	The average of the weights of all incoming of edges from vertex y to vertex x in a network G. Example: $AverageVertexInWeight(G_r,B,D) = -1$
Clustering	Through clustering, communities (set of nodes that is densely connected internally) within the graph can be detected.

Table 1. Graph measures and operations applied to analyze the feedback graphs.

In order to carry out the mapping of this model to a 2-dimensional visualization space, a convenient visualization technique needs to be selected in the subsequent step.

3.1.3 Building Block: Visualization & Interaction Techniques

3.1.3.1 The Reputation Matrix

There is wide range of visualization techniques suitable for network data. Common techniques for displaying graphs can be divided into three groups: node-link, matrix-based and hybrid representation (Landesberger et al., 2011). According to a comparison of node-link and matrix-based representation carried out by Ghoniem et al. (2004), node-link diagrams are particularly suitable for small graphs. Matrices, however, are more convenient for large or dense networks. Using appropriate node ordering, matrices can furthermore easily reveal substructures such as communities in a graph. Common e-commerce environments can involve both a high number of nodes and a large amount of transactions. As we want

to detect collusive behavior by applying different reputation measures, we decided to choose a matrix representation combined with node ordering.

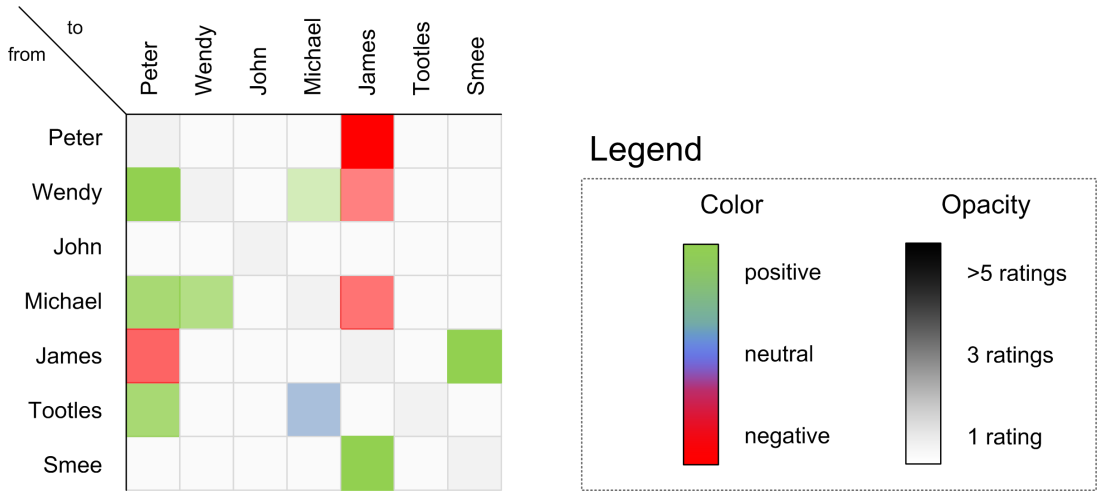

Figure 3. Reputation matrix of the buyer feedback graph for an electronic marketplace with 7 actors and positive, neutral and negative ratings.

In general, a matrix representation visualizes the adjacency matrix of the given graph and displays the edges in the cells. As our graph is not a simple graph but a multigraph with weighted edges, we make use of two further visual levers, namely color and opacity. Analogous to a heat map, where the values contained in a matrix field are represented as colors, we range the colors according to the $AverageVertexInWeight(a,b)$. In this way, we can clearly depict the aggregated ratings from vertex b towards vertex a. To additionally include the strength of a tie (relation) between two nodes, we vary the opacity of a matrix field according to the $VertexInDegree(a,b)$. The result is the reputation matrix as depicted in Figure 3.

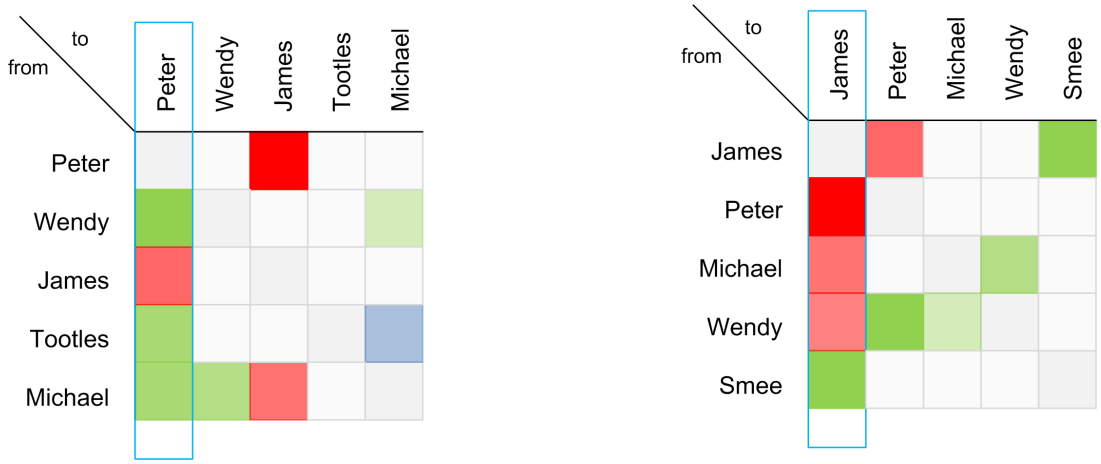

Figure 4. Reduced weighted heat maps for actors "Peter" and "James". All direct ratings toward the actor in focus are listed in the first column.

On most electronic marketplaces, users are not presented an integrated view of the whole feedback network. Instead, they rather focus on the evaluation of a specific seller's feedback profile. Thus, it is not necessary to display the entire feedback data of all actors. Consequently, we reduce the complete matrix to all data relevant for the analysis of the vertex in focus. This again leads to a comparably dense network, where a matrix representation is particularly advantageous. This operation is especially important for large

electronic marketplaces with several thousands up to millions of actors. Visualizing the entire network in a matrix would be impossible because of the limited space of common displays. Figure 4 shows two reduced reputation matrices for the actors "Peter" and "James".

3.1.3.2 Interaction Technique: Order and Filter

A randomly sorted matrix representation such as depicted above does only rarely allow to gain additional insights a metric cannot convey. To identify attacks or anomalies in the feedback data, our matrix needs to be resorted. Vertex ordering, however, is one of the biggest challenges of using a matrix representation (Mueller et al., 2007). An effective way to cope with this limitation is user interaction. Making use of the user's experience and his visual cognitive capabilities, the incorporation of a human analyst and a computer might have an advantage over analyses carried out by machines only. We therefore propose an interaction technique we call "order and filter". Here, the user first applies an ordering mechanism to bring the rows and columns in an order that might allow to reveal correlation and find groups of unfair ratings. Subsequently, he can use a selection tool to select and filter the unfair ratings identified.

Ordering based on	Objective
$AverageVertexInWeight(G_r,x)$, $VertexInDegree(G_r,x)$	**Detection of random unfair ratings.** The *AverageVertexInWeight* and the *VertexInDegree* of the seller graph can be considered to be measures for the recommender reputation. Detecting and discounting false feedback based on recommender reputation has been shown to be an effective mechanism against unfair ratings in several publications (Tavakolifard and Almeroth, 2012). Ordering all actors by recommender reputation in a decreasing manor, unfair ratings would accumulate on the bottom left of the matrix.
$VertexInDegree(G_s,x=a,y)$	**Collusion & Sybil attack detection.** According to Li et al. (2012), collusion is characterized through (among other things) high reputation of all colluding nodes and a high rating frequency attributed to very positive ratings between all colluding parties. Ordering based on the number of feedback relations between node *a* (the actor in focus) and all other related nodes helps us to identify and filter suspicious actors that have provided multiple ratings to artificially increase or decrease node *a*'s reputation. Depending on the ordering (i.e. ascending or descending), we will find a square depicting the relations between the colluding nodes on the top left of the matrix for collusion attacks. Sybil attacks, in contrast, will form a bar.
Clusters	**Collusion & Sybil attack detection.** There are many publications that propose clustering approaches to identify collusion or Sybil attacks in various environments. Clustering is particularly interesting for attacks in which ordering by advisor reputation or number of relations fails. This could be the case if the group of colluding nodes is large enough to gain enough reputation by exchanging only single positive ratings between all nodes. However, applying clustering algorithms to detect communities in the buyer/seller reputation graph and ordering all vertices based on the outcomes of the cluster analysis, these groups can be detected and filtered.

Table 2. Ordering rules to identify malicious behavior that unfairly pushes the reputation of vertex a.

To implement this interaction technique, we set up a range of ordering mechanisms that adapt well-known approaches to discount feedback from the research on trust and reputation systems. Table 2 gives an overview of the ordering rules and the intended objectives. This generic conceptual design can be applied in various application areas. To demonstrate our visualization-based approach, we implement a prototype that can easily be adapted to different rating scales.

3.2 Prototype Implementation

Since most online reputation systems for e-commerce settings are displayed in a browser, we implemented our prototype using a three-tier client-server architecture applying state-of-the-art Web standards. We decided to employ those for an easier integration with an electronic market place in the future. On the data layer, we made use of a hybrid concept combining Couchbase and MySQL as two common solutions for NoSQL and SQL. They contain all data necessary to set up the reputation matrix. Raw data must thereto at least comprise unique users as well as a list of feedback ($feedback\_from$, $feedback\_to$, $rating\_value$). The logic layer was implemented in PHP scripts employing an Apache webserver. The rating interval R of the specific application area can be individually defined by the developer to fit the input data. As our clustering algorithm we chose DBSCAN, which is a "Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" (Ester et al., 1996). As pointed out in Section 2, collusion is characterized by a higher rating frequency as well as a higher density within the colluding nodes. Thus, a density-based standard algorithm suffices our prototypical needs. The case studies introduced in the following section will furthermore demonstrate its effectiveness. The presentation layer was implemented using common Web standards such as HTML5, CSS3, SVG and JavaScript. The visualization in particular was implemented by means of the d3.js-package[5]. D3 (Data-Driven Documents) is a JavaScript library enabling to manipulate documents based on data. When applying D3, data are bound to the browser's DOM. This allows the user to instantly interact and manipulate the visualization. Regarding the implementation of the matrix representation, we referred to an example proposed by Michael Bostock (2012).

4 Empirical Study

In order to demonstrate the proper functioning of our prototype, we introduce two empirical studies of prominent online reputation systems. We use them to show how unfair rating attacks can reliably be identified using our interactive visualization-based approach. The first real-world dataset contains feedback profiles from the electronic marketplace eBay. The second dataset comprises a large number of reviews from the customer review site Epinions.

4.1 Case 1: Electronic Marketplace - eBay

eBay constitutes one of the world's largest electronic marketplace, involving more than 150 million active buyers (eBay, 2014). On eBay, actors can take the role of both buyer and seller. According to a study published by the eBay research lab in 2011, 82.5% of the users only buy while 5.76% of the users only sell (Shen and Sundaresan, 2011). Therefore, only 11.74% of all users take both roles. eBay's reputation system, which can be considered one of the first and most prominent online reputation systems, allows the transaction parties to rate each other after having carried out a transaction. These ratings can take values of the interval $R = \{-1, 0, 1\}$ for $negative$, $neutral$ and $positive$.

Defining our model, we need two graphs – the seller feedback graph G_s and the recommender feedback graph G_r. On eBay, feedback (recommendations) cannot directly be rated. However, buyer reputation (the sum of all ratings from sellers toward a particular buyer) might be a good measure for recommender reputation since transactions cost money, effort and time. Thus, developing buyer reputation is quite expensive. Resnick et al. (2006) point out that in their field experiment on eBay's reputation system, feedback of buyers who had themselves no feedback was not equally considered as feedback from well-established buyers. Therefore, we take feedback created by sellers toward buyers as recommender feedback and feedback created by buyers toward sellers as seller feedback.

[5] http://www.d3js.org/

To demonstrate how unfair ratings can be detected and filtered, we acquired feedback data of several suspicious eBay users. These user profiles could be found through buyer complaints on eBay's community boards. As multiple buyers had complained that they had not received their already payed products, we had reasonable certainty that the respective sellers had malicious intentions. Our example involves a feedback profile of an eBay seller who counts 110 reviews in total created in the last 6 months. 99 of these take a positive, 3 a neutral and only 9 a negative value. This leads to an average of 89% positive ratings in the last 6 months. Figure 5 (left) depicts the reputation matrix for the seller graph of our user in focus. We anonymized the real user names by a sequential number, whereby user 1 is our user in focus.

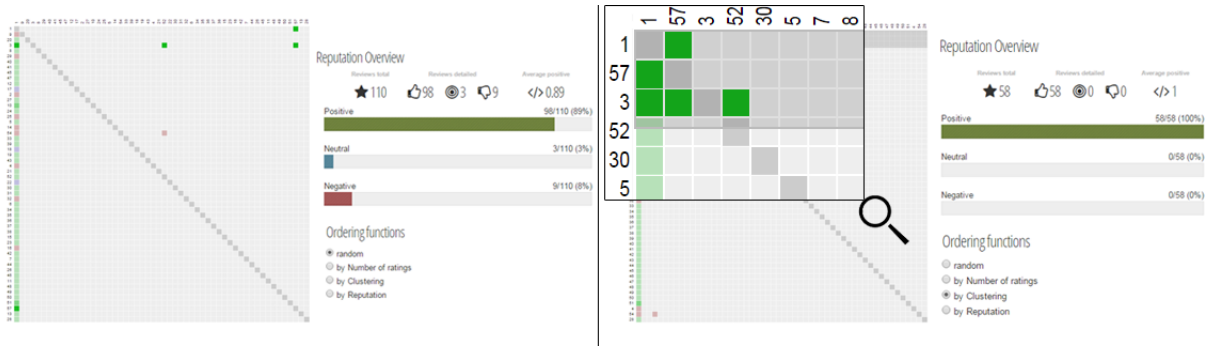

Figure 5. Reputation matrix for the eBay user. Left: Columns are randomly sorted. Right: All feedback but the cluster on the upper left is filtered. (Note: Original screenshot was distorted)

As the reputation matrix is reduced to all users directly related to the user in focus, the first column is completely colored. Obviously, there are nearly no further relations between the other actors. This is presumably due to the fact that most users take the role of buyers. Interactions between actors, however, can only occur if at least one of both parties takes the role of a seller. Furthermore, it is noticeable that there are two users who seem to have rated our user in focus quite often. Reordering the matrix by clusters and selecting the suspicious ratings, we find that more than half of all ratings have been created by two users as depicted in Figure 5 (right). As they mutually rated each other extraordinarily high (compared to the rest), we found collusive behavior. A live demo of this case can be found online[6].

4.2 Case 2: Customer Review Site - Epinions

In our second case study, we turn to the customer review site Epinions, whose main purpose is to enable people to provide feedback on products and services. As these reviews might vary in quality, Epinions furthermore allows its users to rate their helpfulness (see Figure 6). In this way, the trustworthiness of a review/reviewer can be determined. Note that every review can be rated by every user of the platform. As opposed to providing feedback on eBay, there are no transaction costs associated with this process. In our case, we denote the rating of reviews as recommender feedback. We visualize the recommender feedback graph G_r to detect ratings unfairly increasing the reputation of a recommender.

Figure 6. Review process on Epinions.

[6] http://trust.bayforsec.de/unfair_ratings/

For our investigations, we use a large-scale Epinions dataset containing about 132,000 users, 1,560,144 reviews and 13,668,319 ratings (Massa and Avesani, 2006). Ratings on reviews on Epinions are expressed in a five-step scale ranging from *Not helpful* to *Most helpful*, which is internally mapped to values in the interval $R = [1;5]$. The rating distribution of our dataset is as follows: 1: 0.017%, 2: 2.323%, 3: 5.609%, 4: 15.100%, 5: 76.950% (Average: 4.67)[7].
To come up with an illustrative case, we inspected the dataset with respect to three criteria. Firstly, we focused on reviewers whose reviews had at least 100 ratings in total. To be at all able to see collusive behavior, we particularly looked for reviewers for whom a large share of their ratings came from a small number of distinct raters. As a third criterion, the ratings of these suspicious raters should differ significantly from the average ratings. Going through this procedure, we arrived at several interesting cases. Our example visualized in Figure 7 focuses on user 547301 who has written 23 reviews and received 136 ratings for them with an average rating of 4.64.

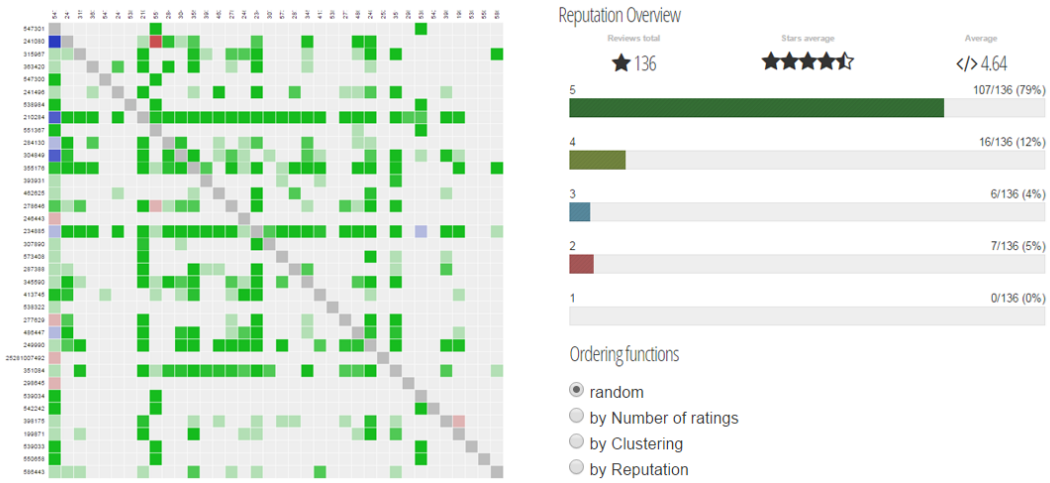

Figure 7. Reputation matrix for the Epinions user. The columns are randomly sorted. (Note: Original screenshot was distorted)

Again, the first column is completely colored because we reduced the reputation matrix to all users directly related to the user in focus. As opposed to Figure 5, we can now observe many interactions between these users. This is presumably due to the fact that almost half of the reviews on Epinions are written by less than 1% of the users (Meyffret et al., 2012). Taking into account that these frequent reviewers are likely to be frequent raters as well, the chances are high that some of them are included in any reputation matrix. Patil et al. (2013) even identified an overlap of around 25% between the top-1% reviewers and top-1% raters. Reordering this matrix by clusters and analyzing the large cluster at the bottom, we can confirm these findings (see Figure 8).
As a consequence, the small cluster on the upper left side could depict collusive behavior. Selecting the bottom cluster (filtering the upper cluster), we found that 73% of the total ratings were created by the upper cluster. All of them carry the maximum positive value. Filtering these ratings reveals a completely different picture with an average rating of just 3.68. Considering an average of 4.67 of all ratings in our dataset, user 547301 comes off comparably badly. Ordering the nodes by reputation (exogenous discounting), we furthermore found that most of the users involved in the upper cluster had a worse reputation than the rest. Here, we can distinguish between colluding nodes (539034, 551367), who could also profit from the collusive behavior, and Sybil nodes (539033, 547300, 550658, 542242, 538984), who could not gain any advantage for themselves. A live demo of this case has been made publicly available[8].

[7] As suggested by the publishers of the dataset, the special rating value 6 is treated as 5.
[8] http://trust.bayforsec.de/unfair_ratings/

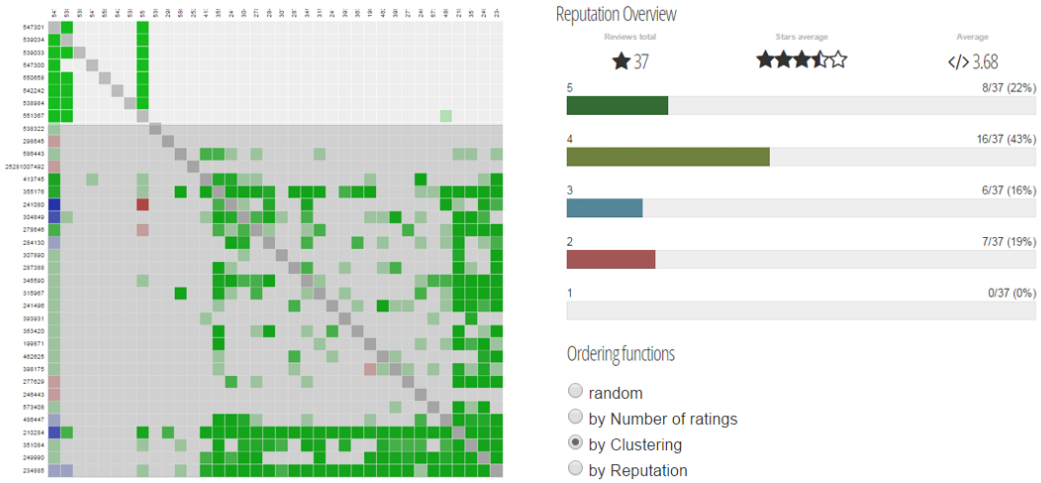

Figure 8. Reputation matrix for the Epinions user. The cluster on the bottom right is selected and other ratings are then filtered. (Note: Original screenshot was distorted)

Overall, these empirical studies demonstrate that an interactive visualization can notably increase the transparency of a reputation system by depicting all input data in one picture. Using our prototype, we are able to show that unfair ratings – particularly in the form of collusive behavior and Sybil attacks – can reliably be detected and filtered. Through the incorporation of the human analyst and the interactive switching between endogenous and exogenous discounting methods, we allow to make well-informed and comprehensible decisions.

5 Conclusion

In this paper, we made a step toward applying visual analytics to detect unfair ratings in online trust and reputation systems. We first introduced a generic conceptual design of a visual analytics methodology that can easily be adapted to various use cases. Subsequently, we demonstrated how this generic concept can be implemented in a prototype for the example of an e-commerce scenario. The empirical study based on real-world data from eBay and Epinions revealed a variety of benefits of our interactive approach. However, there are some shortcomings that could be topic of our future work.

On the one hand, our reduced matrix might not reveal advanced attacks that go beyond the direct interaction between the actor in focus and his colluding partners. This in turn might be tolerable since advanced attacks require enormous financial investments in most cases. On the other hand, the ordering metrics employed in our prototype such as the reputation measure are still quite simple. An improvement of these in future work might also support the detection of more sophisticated attacks. Moreover, for converting our prototype into a final product, introducing multi-assignment clustering algorithms could reveal multiple collusive clusters. Another fruitful enhancement could involve additional interaction techniques that tap the full mindset of the human analyst.

Throughout this work, we showed that interactive visualizations bare an enormous potential that could lead to substantially more robust reputation systems and enhanced user experience. Integrating our methodology into an e-commerce environment, organizations could be aided in increasing the overall transparency of their reputation systems.

6 Acknowledgements

The research leading to these results was supported by the “Bavarian State Ministry of Education, Science and the Arts” as part of the FORSEC research association.

References

Artz, D. and Y. Gil (2007). "A Survey of Trust in Computer Science and the Semantic Web." *Web Semantics: Science, Services and Agents on the World Wide Web* 5 (2), 58–71.

Bostock, M. (2012). *Les Misérables Co-occurrence*. URL: `http://bost.ocks.org/mike/miserables/` (visited on 11/27/2014).

Cornelli, F. et al. (2002). "Choosing Reputable Servents in a P2P Network." In: *Proc. of the 11th International Conference on World Wide Web (WWW)*, pp. 376–386.

Dellarocas, C. (2000). "Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior." In: *Proc. of the 2nd ACM Conference on Electronic Commerce (EC)*, pp. 150–157.

— (2006). "Reputation Mechanisms." In: *Handbook on Economics and Information Systems*. Ed. by T. Hendershott. Elsevier Publishing, pp. 629–660.

Diekmann, A. et al. (2014). "Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets." *American Sociological Review* 79 (1), 65–85.

eBay (2014). *eBay Marketplaces Fast Facts At-A-Glance (Q3 2014)*. URL: `http://www.ebayinc.com/sites/default/files/MP_FastFacts_Q32014.pdf` (visited on 11/27/2014).

Ester, M. et al. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In: *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 226–231.

Gambetta, D. (1988). *Trust: Making and Breaking Cooperative Relations*. New York, NY, USA: Blackwell.

Ghoniem, M. et al. (2004). "A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations." In: *Proc. of the IEEE Symposium on Information Visualization (INFOVIS)*, pp. 17–24.

Gupta, M. et al. (2003). "A Reputation System for Peer-to-Peer Networks." In: *Proc. of the 13th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, pp. 144–152.

Habib, S. M. et al. (2013). "A Trust-Aware Framework for Evaluating Security Controls of Service Providers in Cloud Marketplaces." In: *Proc. of the 12th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 459–468.

Hammer, S. et al. (2013). "A User-Centric Study Of Reputation Metrics in Online Communities." In: *Proc. of the 3rd Workshop on Trust, Reputation and User Modeling (TRUM)*.

Hevner, A. R. et al. (2004). "Design Science in Information Systems Research." *MIS Quarterly* 28 (1), 75–105.

Jøsang, A. and J. Golbeck (2009). "Challenges for Robust Trust and Reputation Systems." In: *Proc. of the 5th International Workshop on Security and Trust Management (STM)*.

Jøsang, A. and R. Ismail (2002). "The Beta Reputation System." In: *Proc. of the 15th Bled Conference on Electronic Commerce*, pp. 41–55.

Keim, D. A. et al. (2010). *Mastering the Information Age - Solving Problems with Visual Analytics*. Goslar, Germany: Eurographics Association.

Landesberger, T. v. et al. (2011). "Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges." *Computer Graphics Forum* 30 (6), 1719–1749.

Li, Z. et al. (2012). "Collusion Detection in Reputation Systems for Peer-to-Peer Networks." In: *Proc. of the 41st International Conference on Parallel Processing (ICPP)*, pp. 98–107.

Littlestone, N. and M. K. Warmuth (1994). "The Weighted Majority Algorithm." *Information and Computation* 108 (2), 212–261.

Massa, P. and P. Avesani (2006). "Trust-aware Bootstrapping of Recommender Systems." In: *Proc. of the ECAI Workshop on Recommender Systems*, pp. 29–33.

Meyffret, S. et al. (2012). *RED: A Rich Epinions Dataset for Recommender Systems*. Université de Lyon.

Mueller, C. et al. (2007). "A Comparison of Vertex Ordering Algorithms for Large Graph Visualization." In: *Proc. of the 6th International Asia-Pacific Symposium on Visualization (APVI)*, pp. 141–148.

O'Donovan, J. et al. (2007). "Extracting and Visualizing Trust Relationships from Online Auction Feedback Comments." In: *Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2826–2831.

Patil, A. et al. (2013). "Quantifying Social Influence in Epinions." In: *Proc. of the 2013 International Conference on Social Computing (SocialCom)*, pp. 87–92.

Peffers, K. et al. (2007). "A Design Science Research Methodology for Information Systems Research." *Journal of Management Information Systems* 24 (3), 45–77.

Resnick, P. and R. Zeckhauser (2002). "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System." *Advances in Applied Microeconomics* 11, 127–157.

Resnick, P., R. Zeckhauser, et al. (2006). "The Value of Reputation on eBay: A Controlled Experiment." *Experimental Economics* 9 (2), 79–101.

Sänger, J. and G. Pernul (2014a). "Reusability for Trust and Reputation Systems." In: *Trust Management VIII: Proc. of the 8th IFIP WG 11.11 International Conference (IFIPTM)*, pp. 28–43.

— (2014b). "Visualizing Transaction Context in Trust and Reputation Systems." In: *Proc. of the 9th International Conference on Availability, Reliability and Security (ARES)*.

Shen, Z. and N. Sundaresan (2011). "eBay: An E-commerce Marketplace As a Complex Network." In: *Proc. of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 655–664.

Sun, Y. L. and Y. Liu (2012). "Security of Online Reputation Systems: The Evolution of Attacks and Defenses." *IEEE Signal Processing Magazine* 29 (2), 87–97.

Tavakolifard, M. and K. C. Almeroth (2012). "A Taxonomy to Express Open Challenges in Trust and Reputation Systems." *Journal of Communications* 7 (7), 538–551.

TripeAdvisor (2012). *How to Remove Defamatory Reviews from TripAdvisor*. URL: `http://www.tripeadvisor.com` (visited on 11/27/2014).

Wang, W. and A. Lu (2006). "Visualization Assisted Detection of Sybil Attacks in Wireless Networks." In: *Proc. of the 3rd International Workshop on Visualization for Computer Security (VizSEC)*, pp. 51–60.

Whitby, A. et al. (2004). "Filtering Out Unfair Ratings in Bayesian Reputation Systems." In: *Proc. of the 7th International Workshop on Trust in Agent Societies*.

Winkler, T. et al. (2007). "Trust Indicator Modeling for a Reputation Service in Virtual Organizations." In: *Proc. of the 15th European Conference on Information Systems (ECIS)*, pp. 1584–1595.

Yang, Y. et al. (2009). "Defending Online Reputation Systems Against Collaborative Unfair Raters through Signal Modeling and Trust." In: *Proc. of the 2009 ACM Symposium on Applied Computing (SAC)*, pp. 1308–1315.

Yu, B. and M. P. Singh (2003). "Detecting Deception in Reputation Management." In: *Proc. of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 73–80.

Zhang, J. et al. (2008). "A Detailed Comparison of Probabilistic Approaches for Coping with Unfair Ratings in Trust and Reputation Systems." In: *Proc. of the 6th Annual Conference on Privacy, Security and Trust (PST)*, pp. 189–200.