Reputation Systems Evaluation Survey

ELENI KOUTROULI and APHRODITE TSALGATIDOU, National and Kapodistrian University of Athens

Various reputation systems have been proposed for a broad range of distributed applications, such as peer-to-peer, ad-hoc, and multiagent systems. Their evaluation has been mostly based on proprietary methods due to the lack of widely acceptable evaluation measures and methodologies. Differentiating factors in various evaluation approaches include the evaluation metrics, the consideration of the dynamic behavior of peers, the use of social networks, or the study of resilience to specific threat scenarios. The lack of a generally accepted common evaluation framework hinders the objective evaluation and comparison of different reputation systems. Aiming at narrowing the gap in the research area of objective evaluation of reputation systems, in this article, we study the various approaches to evaluating and comparing reputation systems, present them in a taxonomy, and analyze their strengths and limitations, with special focus on works suggesting a Common Evaluation Framework (CEF). We identify the challenges for a widely accepted CEF that enables testing and benchmarking of reputation systems, and we present the required properties for such a CEF; we also present an analysis of current CEF-related works in the context of the identified properties and our related proposals.

Categories and Subject Descriptors: C.2.0 [**Computer-Communication Networks**]: General—*Security and protection*; C.4 [**Computer Systems Organization**]: Performance of Systems—*Reliability, availability, serviceability*; H.4.3 [**Information Systems Applications**]: Communications Applications; I.6.6 [**Simulation and Modeling**]: Simulation Output Analysis; I.6.7 [**Simulation and Modeling**]: Simulation Support Systems; K.4.4 [**Computers and Society**]: Electronic Commerce—*Security*

General Terms: Reliability, Security, Design, Performance

Additional Key Words and Phrases: Reputation systems, reputation system evaluation, benchmarking, simulation, reputation systems comparison, reputation systems attacks, threat analysis, credibility, commonly agreed evaluation framework

ACM Reference Format:
Eleni Koutrouli and Aphrodite Tsalgatidou. 2015. Reputation systems evaluation survey. ACM Comput. Surv. 48, 3, Article 35 (December 2015), 28 pages.
DOI: http://dx.doi.org/10.1145/2835373

1. INTRODUCTION

The area of reputation systems has received an increasing level of attention from the research community in the past years. Various reputation systems have been proposed for a broad range of distributed applications, such as peer-to-peer (P2P) [Zhang and Wang 2012; Huang et al. 2006; Wang et al. 2006; Marti and Garcia-Molina 2006; Kamvar et al. 2003; Aberer and Despotovic 2001], ad-hoc [Mármol and Pérez 2012; Buchegger and Le Boudec 2004; Almenarez et al. 2004], and multiagent (social network–based)

Authors' addresses: E. Koutrouli and A. Tsalgatidou, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens; emails: {ekou, atsalga}@di.uoa.gr.

DOI: http://dx.doi.org/10.1145/2835373

applications [Serrano et al. 2012; Yu and Singh 2003; Sabater and Sierra 2001]. The common characteristic of those applications is that different entities offer services to each other, whereas the goal of a reputation system is to help an entity in a distributed community choose a trusted entity to interact with. This makes the role of a reputation system vital for the success of such applications. Fulfilling this role is not a trivial task: To be effective, the design of a reputation system for an e-community should take into consideration several application–related aspects and possible threats. Therefore, the objective and adequate evaluation of reputation systems by reputation system designers is essential to their success. The availability of a common framework and tools for simulation and evaluation of reputation systems can thus provide valuable help in making the right choices when designing a reputation system for a particular application environment.

Up to now, the evaluation of reputation systems has been mostly based on proprietary methods due to the lack of widely acceptable evaluation measures and methodologies. As a consequence, reputation systems comparison is not straightforward. These challenges in objective evaluation and comparison of reputation systems hinder the easy assessment of the quality of a reputation system, its fine-tuning, and also decisions about which reputation system an application developer should choose for integration into her applications. One of the reasons for this situation is the lack of standardization in reputation system design, which results in lack of interoperability between reputation models for distributed systems; this makes the task of establishing common evaluation measures for reputation systems even more difficult [Mármol and Pérez 2010]. Differentiating factors in the various simulation and evaluation approaches stem from the diverse application context of reputation systems and the specific goals of reputation system designers. These factors include the incorporation of the dynamic behavior of peers (e.g., their ability to enter and leave the system freely, which is considered in some works—Zhou and Hwang [2007] and Liang and Shi [2005]—and not considered in others), different evaluation measures, the use of social relationship information (e.g., acquaintances in Walsh and Sirer [2005, 2006] or coordination/competition relationships in Sabater and Sierra [2001]), or the study of resilience in specific threat scenarios.

Nevertheless, some research works try to establish a common evaluation framework for reputation systems. They choose either a theoretic criteria framework (e.g., Ruohomaa et al. [2007] and Noorian and Ulieru [2010]), a simulation-based approach (e.g., West et al. [2010] and Mármol and Pérez [2009]), or a combination of both theoretic and simulation-based approaches (e.g., Hazard and Singh [2009, 2013]). They also use different evaluation metrics and specific threat scenarios. The ability of these works to be used as general frameworks for different cases of reputation systems has limitations because some of them refer to specific components of reputation systems (e.g., only to reputation metrics [Schlosser et al. 2006]) or enable the evaluation of specific types of reputation systems (e.g., either centralized as in Celestini et al. [2013a, 2013b] or decentralized as in Zhang et al. [2007]). A number of challenges are related to the establishment of a commonly accepted and generally used evaluation framework, including the lack of standardization of reputation systems, the lack of agreement on the evaluation metrics that will be used, specific idiosyncrasies of the environment leveraged by reputation systems that make their comparison difficult, or the ability to incorporate specific threat scenarios.

Motivated by (i) the need for the objective evaluation and comparison of different reputation systems that would be enabled by a commonly accepted evaluation framework for reputation systems and (ii) the lack of such a generic evaluation framework and the limitations of current related works, we envision the need for a Common Evaluation Framework (CEF) that can address the current challenges. We believe that most reputation system researchers have come across this requirement when faced

with decisions about how to properly evaluate their system. Our motivating scenario consists of various cases in which a reputation system designer is required to:

—evaluate a reputation system that he or she has designed for a specific application,
—fine-tune this reputation system, or
—compare different reputation systems in a specific application context or threat scenario in order to select the right reputation system among them.

To address these requirements, we conducted research in the area of reputation systems evaluation in order to present the various categories of approaches used, to identify and evaluate frameworks that can be used to easily simulate reputation systems and their responses to most related attacks, and to determine which could provide reliable and commonly agreed-on evaluation metrics.

More specifically, our work has two main objectives: The first is to provide a taxonomy of current works on the evaluation of reputation systems derived from a thorough review of the research area. The goal of this taxonomy is to help reputation system designers make decisions regarding devising and implementing a suitable evaluation method for their system, using a suitable available framework for evaluating and fine-tuning their system, and comparing different reputation systems.

The second objective of our work is to define the basic requirements and characteristics of a commonly accepted simulation framework that can be used for the objective evaluation and comparison of reputation systems.

To attain our objectives, we proceeded as follows:

—we thoroughly studied approaches to reputation systems evaluation,
—we classified these approaches according to their spectrum of usage and to their nature; then, we analyzed those categories with a focus on works on CEFs that aim at enabling objective evaluation and reasonable comparison of different reputation systems,
—we presented research works on CEFs according to our taxonomy,
—we derived the limitations of current simulation-based CEFs and defined the desirable characteristics of a generally accepted CEF,
—we analyzed the current simulation-based CEFs according to the aforementioned characteristics, and
—we finally derived a set of factors that a CEF should have, aiming at leading toward the definition of improved CEFs.

The results of our work are presented in the remainder of this article, which is organized as follows: The second section presents current approaches to the evaluation of reputation systems found in the literature, organized in a taxonomy comprising custom-made approaches and proposals for common evaluation and comparison frameworks. The various categories are thoroughly analyzed with a focus on works on CEFs. The third section presents our work toward fulfilling the second objective; specifically, it discusses the limitations of current works on the definition of a common framework that can objectively evaluate most reputation systems in various application environments. It also presents the desirable properties of a commonly agreed-on framework for evaluation and benchmarking. This is followed by an analysis of current research works on simulation-based CEFs according to these properties and related proposals for the design of a CEF that would satisfy these properties. In the fourth section, we present our concluding remarks.

2. EVALUATION APPROACHES: A TAXONOMY

Due to the lack of a commonly agreed-on evaluation framework, reputation system designers try to fulfil the need for reputation system evaluation by making their own

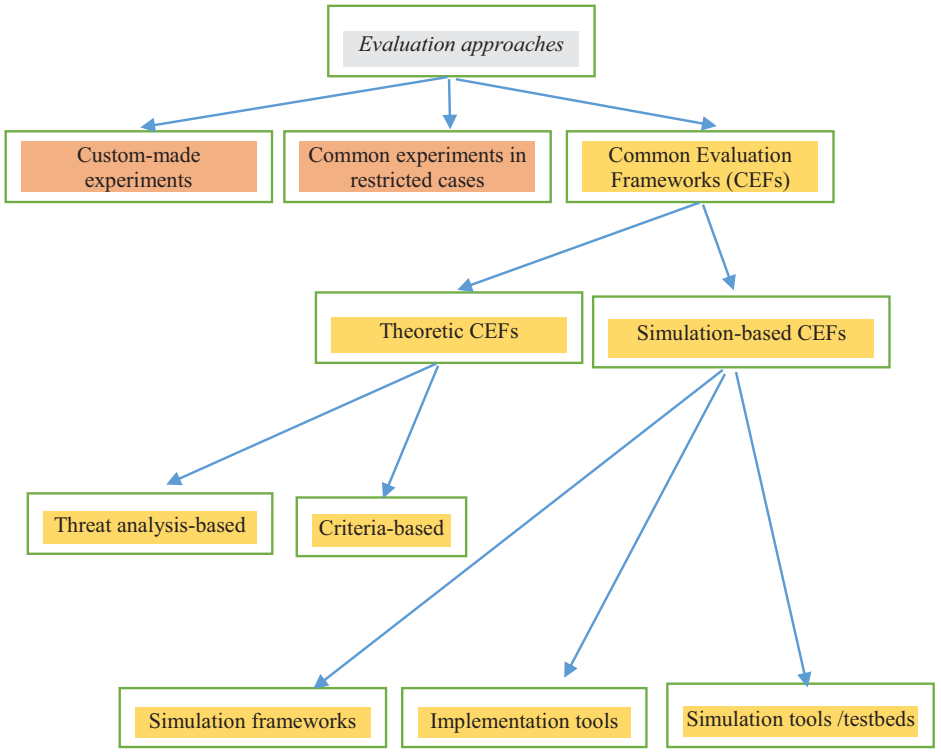

Fig. 1. Taxonomy of evaluation approaches.

decisions regarding the approach they will choose. They can choose to use either custom-made simulation experiments and evaluation criteria, experiments that have been designed and used by other researchers, or some available tools or frameworks that have been designed with the goal of facilitating simulation and objective evaluation.

Individual reputation systems proposed in the literature have been mainly evaluated with the use of custom-made simulators and custom experiments. Furthermore, some research works have proposed evaluation mechanisms for the comparison of different reputation systems under a specific set of scenarios, which cannot be applied directly to all cases. Such mechanisms include experiments based on the Prisoner's Dilemma (PD) game, as well as experiments that specify specific test scenarios and have been used for the evaluation and comparison of specific reputation systems for e-commerce environments. These common experiments have been explicitly designed for specific scenarios shared among a number of reputation system. They cannot, however, be considered generic evaluation approaches for the evaluation and comparison of heterogeneous reputation systems. The need for generic evaluation approaches led to a number of research works that focus on the development and use of generic frameworks for evaluation and comparison of reputation systems; we refer to these as CEFs. These works are either theoretic (i.e., they study how a reputation system deals with a number of criteria or attacks), or they offer simulation and implementation platforms/tools for the evaluation, comparison, and fine-tuning of reputation systems through experimentation.

The various available approaches for reputation systems evaluation can be classified according to the taxonomy presented in Figure 1. In the remainder of this section, we present this taxonomy: We first analyze those works that use custom-made simulation

experiments or experiments designed by other researchers (i.e., common experiments that apply in restricted cases; see Section 2.1). In the analysis of individual works, we chose a number of reputation systems in various application areas, such as P2P systems, participatory sensing systems, and ad-hoc networks, and we examine the evaluation approaches followed by their designers regarding the examined population, the evaluation metrics used, the kind of threat scenarios checked, and also whether the dynamic behavior of entities is incorporated in the simulation or not. We also present some common experiments for evaluation (see Section 2.2). We discuss the limitations of the two analyzed categories of evaluation approaches, which create the need for more generic evaluation frameworks, and then we continue with a thorough presentation of works related to CEFs according to the taxonomy presented in Figure 1 (Section 2.3). Special focus is given on simulation-based CEFs (Section 2.3.2) that offer more possibilities for objective evaluation and comparison than do theoretic frameworks.

2.1. Reputation Systems Evaluation Based on Custom-Made Experiments

Works that present individual reputation systems evaluate them in a proprietary manner usually based on simulation. Simulation experiments differ among various reputation system works. They differ both in the simulation parameters (e.g., the simulated population) and in nature (e.g., in the properties that are tested and the simulated scenarios). Experiments focus either on the *correctness of predictions* (through the difference between the estimated reputation and the actual reputation values or the rate of successful transactions) or on the *performance* of the system (e.g., the time needed for the system to converge to a reputation value), and there are no benchmarks for either.

Table I shows the differences in simulations of a number of representative reputation systems through

(1) the evaluated properties,
(2) the simulated network size,
(3) the reputation system threats considered in the evaluation of each reputation system, and
(4) whether the dynamic nature of entities is simulated or not.

Regarding the evaluated properties, we present the various evaluation metrics used, which can be attributed to two main properties: the *correctness of predictions* and/or the *performance* of the systems.

We also examine if the dynamic behavior of peers (e.g., their intermittent presence in the system or topology changes) is simulated and the kind of threat analysis involved, which usually considers how a reputation system behaves in one or more of the following attacks:

—*badmouthing*, when peers send negative recommendations regarding honest peers,
—*collusive manipulating of reputation*, when peers form collusions to boost their own reputation with positive recommendations for each other and decrease the reputation of honest peers by sending unfairly negative recommendations regarding them,
—*oscillating behavior*, when peers have an unstable transactional behavior, trying to keep high reputation values while cheating in a small fraction of transactions, and
—*traitors attack*, when peers behave honestly in their transactions for some time to acquire a high reputation value and then start cheating.

These types of attacks, which are only a subset of a variety of attacks and misbehavior types that can occur in reputation systems [Koutrouli and Tsalgatidou 2012], are simulated with different experiments in some of the studied reputation systems.

We chose for our analysis 15 prominent reputation systems from a plethora of research works in various application contexts, with the aim of exhibiting their diverse evaluation approaches. A short description of these systems is given to point out how the individual evaluation approaches have been based on the specifics of the studied systems. First, we present 10 reputation systems for P2P communities categorized with respect to their goal, the scope of reputation (global or local, depending on whether it is based on global information or determined locally through peers using personalized information), the reputation estimation algorithm, and their evaluation approaches. Then, we similarly present five reputation systems for mobile or ubiquitous applications, featuring their application-specific requirements and their special reputation estimation and evaluation approaches.

Aberer et al. [2001], PeerTrust [Xiong and Liu 2004], FuzzyTrust [Song et al. 2005], PowerTrust [Zhou and Hwang 2007], and EigenTrust [Kamvar et al. 2003] estimate global reputation values for entities in P2P communities. The first three of these systems assume self-organized structures for the organization of peers, whereas the latter two use super peers for reputation estimation. Regarding the reputation estimation algorithm, Aberer et al. [2001] and PeerTrust aggregate transaction-based feedbacks, whereas the others estimate global reputation values by aggregating local scores that have been estimated by peers based on their direct transactions. PeerTrust and FuzzyTrust also integrate context characteristics such as quantity and price in the estimation of reputation values.

Different experiments and evaluation metrics are used in these global reputation systems. Aberer et al. [2001] evaluate the *reliability of trust assessments* as a ratio of unsuccessful trust assessments, including both cases that led to a decision for transaction and those that did not lead to such a decision and giving more weight to the first cases. They experiment with different population sizes (128 vs. 1,024 nodes) and with various values for the parameters of the *cheater population*, the *number of interactions,* and *the probability of peers to lie*. A related evaluation metric, the *success rate of transactions,* is used as an evaluation metric in PeerTrust [Xiong and Liu 2004] and EigenTrust [Kamvar et al. 2003]. More specifically, the *percentage of unsuccessful transactions* is estimated in Kamvar et al. [2003] in a number of threat scenarios, whereas the *transaction success rate* is estimated in PeerTrust [Xiong and Liu 2004], together with the *computation error*, which is the Root-Mean-Square (RMS) of the computed and actual trust values of all peers. The authors also evaluate variations of the PeerTrust reputation metric regarding their effectiveness against oscillatory behavior. The RMS error is also used in PowerTrust, together with the number of simulation iterations needed to converge to a reputation value considered actual (*convergence overhead*) and the difference between estimated and actual reputation value (*ranking discrepancy*). A similar metric to the *convergence overhead* of PowerTrust is the *convergence time*—the *time needed* to estimate the global reputation for each peer—which is used in Song et al. [2005], together with the *detection rate* of malicious peers and the *message overhead* involved in global reputation aggregation. Song et al. [2005] use for their experiments a varying number of peers and trace data from the eBay system [eBay 2014]; they compare their system with EigenTrust [Kamvar et al. 2003].

NICE [Sherwood et al. 2006], PET [Liang and Shi 2005], Marti and Garcia-Molina [2004], Yu and Singh [2002], and Credence [Walsh and Sirer 2005, 2006] estimate local reputation values of entities in P2P e-communities. PET [Liang and Shi 2005], Marti and Garcia-Molina [2004], and Credence [Walsh and Sirer 2005, 2006] describe reputation systems in a file-sharing context, whereas Yu and Singh [2002] and NICE [Sherwood et al. 2006] present reputation algorithms for open societies in a general context, using recommendation chains.

In NICE [Sherwood et al. 2006], peers aggregate feedback from other peers about their own behavior in transactions and store it themselves. The authors evaluate their reputation system through simulations that explore its specific parameters (e.g., the amount of information stored in each peer and the simulated system sizes, which range from 512 to 2,048 users). They analyze the system's *scalability* in terms of *storage* and *run-time costs* and the *robustness* of the system in terms of *how many transactions it takes for the system to stabilize*, and they experiment with a simple malicious user model. A similar metric, the *reputation convergence* in terms of the number of reputation queries needed for the average reputation value to stabilize, is used by Yu and Singh [2002], who present a reputation algorithm based on direct and indirect information from third parties. Recommendation chains are formed between trusted peers and experiments are made for examining the *bootstrapping phase* (e.g., what happens when simulations starts), the *traitors attack*, and the effect of the *percentage of noncooperative agents* to reputation convergence. The simulated population size ranges from 100 to 500 agents to study how the system scales.

The PET reputation model [Liang and Shi 2005] integrates a risk factor in the reputation estimation to effectively adapt to oscillatory behavior of peers. The simulation uses an economic model for resource payments and simulates 500 servers and a number (4,700 or 9,400) of clients. The authors define and estimate four metrics in their evaluation experiments: *sensitiveness* (number of peers with whom a peer has had direct transactions), *hit ratio* (the number of honest peers that a peer knows), *effectiveness* (average number of transactions needed to find a specific number of good entities), and *applicability* (percentage of successful transactions), and they explore the effect of various risk- and reputation-related parameters. Marti and Garcia-Molina [2004] make several experiments for evaluation and fine-tuning of their proposed reputation system for P2P file-sharing applications using a custom-made P2P simulator. They measure efficiency through a *verification ratio metric*, which is the ratio of resource verification checks in relation to successful queries, and also through the *load* and *message traffic* generated. Peer presence in the system is not static because a peer can be replaced by another peer. Finally, Walsh and Sirer [2005, 2006] propose Credence, which is a reputation system for objects (files) shared in a P2P file-sharing application. Credence is based on peers' recommendations about the authenticity of an object, and these are weighted according to the similarity of peers regarding their recommendations on a common set of objects. For the simulation of Credence, the authors in Walsh and Sirer [2006] used shared files from a set of 681 Credence clients, whereas in Walsh and Sirer [2005] they used experiments with up to 2,000 clients to evaluate *convergence* as (i) the *success rate over time* and (ii) the *size of correlation table over time* and *accuracy* as the *size of correlation table* of each peer. Different situations are examined, including scaling, dynamic behavior, and several attacks.

A new generation of reputation systems has more recently emerged to support new types of applications that are based on mobile communications, such as pervasive social chatting [Yan et al. 2013], participatory mobile sensing [Huang et al. 2012; Wang et al. 2013], object tracking applications [Roosta et al. 2006], and Vehicular Ad Hoc Networks (VANETs) [Mármol and Pérez 2012]. In these applications, services are exchanged between entities which are connected through wireless networks via portable devices or sensors and are usually unknown to each other. The goal of these reputation systems is to dynamically estimate the trustworthiness of a specific target, which can be either an entity (e.g., a service provider or Trusted Third Party [TTP]) or an object (e.g., data item). Reputation is usually estimated either globally by an application server or locally by each node. Special focus is given to privacy, energy conservation, and secure communications. Consequently, the evaluation of these systems integrates in the examined scenarios attacks related to the identification of nodes, such as

privacy breaching, as in the case of revealing the real IDs of nodes where anonymity is required, or *impersonation*, in which a node participates in the system using a user ID belonging to someone else. Furthermore, *mobility* is usually considered in the simulations. The evaluation of energy consumption is related to the length of the *routing path* or the *convergence speed*, which are estimated in some cases. We included the referenced reputation systems in our analysis and describe them shortly in the following paragraphs.

Huang et al. [2012] propose a global reputation system for participatory sensing applications in which users are requested to collect sensor data and to upload them to an application server. Their focus is in preserving the anonymity of users through the use of suitable anonymization techniques. A user reputation value is calculated by the application server using ratings resulting from multiple interactions with the user. The authors evaluate their system using a real-world participatory sensing application and a set of 150 reporting devices. For their evaluation, they estimate (i) the *accuracy* through the *round square mean error* resulting from the reported data and the actual data and (ii) the "*linkability*"; that is, the likelihood of linking contributions from the same user over time and thus identifying the contributing user. Wang et al. [2013] also propose an anonymous reputation system for participatory sensing applications that estimates user reputation and uses it together with context factors, such as time and location, for the estimation of data trustworthiness. They simulate the system with 100 nodes per task and evaluate *rates of false-positive* and *false-negative data* and also *how quickly the reputation of a malicious user converges* to its real value in the case of various rates of malicious users who intentionally send false data.

Roosta et al. [2006] present a reputation system for object tracking applications in which nodes (sensors) communicate information regarding the location of objects; nodes also estimate local reputation values for their neighbors using beta distributions. The attack scenario considered involves malicious users that compromise other nodes physically in order to participate in the system and inject faulty data. The evaluation of the reputation system is done using a grid of 2,500 nodes and estimating the *average error* of the estimated number of tracks compared with the actual number of tracks.

Mármol and Pérez [2012] propose a reputation system for VANETs, in which vehicles communicate wirelessly and exchange messages regarding road conditions. Each node computes a reputation value for other nodes based on its previous reputation estimation for the same node and the recommendations received by its neighbors, weighted with the recommenders' reliability and the recommendation received by special trusted units of the infrastructure. The system is simulated and evaluated by using 50–150 simulated nodes and by examining the *rate of selected trustworthy nodes* in relation to varied numbers of malicious users who spread false messages and/or unfair recommendations.

Yan et al. [2013] propose a reputation system for pervasive social chatting using mobile ad-hoc networks. Special trusted servers manage node real IDs and pseudonyms and compute global reputation. Then, nodes estimate local reputation values for other nodes based on (i) global reputation values, (ii) votes for messages, and (iii) deviation of opinions. For the evaluation of the system, the evolution of local reputation values (*reputation convergence*) is examined in relation to the number of exchanged messages in different scenarios. A prototype-based user study is also used for assessing the usability of the reputation system.

From the description of the studied systems and the analysis of Table I, we notice that, although some evaluation properties/metrics are similar in nature, they are estimated differently; for example, the *convergence* measures, which refer either to the number of transactions needed for reputation values to stabilize, as in Yu and Singh [2002] and NICE; to the success rate over time, as in Walsh and Sirer [2005]; or to the time needed to establish a global reputation value, as in FuzzyTrust [Song et al. 2005].

Table I. Analysis of Evaluation Approaches of Individual Reputation Systems

Reputation System	Simulated Network Size	Properties		Threat Scenarios/Dynamic behavior
		Performance	Correctness of Prediction	
NICE Sherwood et al. [2006]	512–2,048	Storage, runtime costs, # transactions for convergence	–	Collusive badmouthing
Aberer et al. [2001]	128–1,024	–	Ratio of the unsuccessful trust assessments	–
FuzzyTrust Song et al. [2005]	100–10,000	Convergence time, message overhead	Ratio of detected malicious peers	–
PeerTrust Xiong and Liu [2004]	Default 128	–	Computation error, success rate of transactions	Oscillatory behavior, dynamic behavior
PowerTrust Zhou and Hwang [2007]	1,000	Convergence overhead	Computation error, ranking discrepancy	Badmouthing, collusion, dynamic behavior
PET Liang and Shi [2005]	500	Sensitiveness, hit ratio, effectiveness	Applicability	Badmouthing, dynamic behavior
EigenTrust Kamvar et al. [2003]	20–105	–	Percentage of inauthentic files	Collusion, oscillatory behavior, dynamic behavior
Marti and Garcia-Molina [2004]	1,000	Load (# verification checks/# queries), message traffic	Verification ratio	Badmouthing, collusion, dynamic behavior
Yu and Singh [2002]	100–500	# Transactions for convergence	–	Oscillating behavior, traitors, dynamic behavior
Credence Walsh and Sirer [2005, 2006]	681–1,000	Success rate over time and size of correlation table over time		Badmouthing, dynamic behavior
Huang et al. [2012]	150	"Linkability" (the likelihood of linking contributions from the same user over time; i.e., identifying the contributing user)	Round Mean Square Error	Dynamic behavior through mobility, badmouthing, privacy breaching
Roosta et al. [2006]	2,500	Average error	–	Dynamic behavior, impersonation
Wang et al. [2013]	100	Reputation convergence	false positive and false negative rates	Collusive misbehavior
TRIP [Mármol and Pérez 2012]	50–150	Rate of selected trustworthy nodes when rates of malicious nodes vary		Collusions where malicious users send false messages and/or unfair recommendations
Yan et al. [2013]	5	Reputation convergence	–	Collaborative badmouthing

Some evaluation measures are devised by the authors of individual systems for testing particular cases. The experiments also differ regarding the scenarios tested. Thus, the heterogeneity of proprietary evaluations does not allow for the comparison of different reputation systems.

2.2. Common Evaluation Experiments under Restricted Scenarios

The evaluation works in this category use a common set of experiments for the evaluation and comparison of different reputation systems in restricted scenarios (e.g., for evaluating reputation systems targeting a specific application environment or utility-based reputation systems). Therefore, they cannot be considered general evaluation frameworks. For example, simulations that assume binary representation of trustworthiness may not adequately evaluate reputation systems with more expressive representations of trustworthiness.

2.2.1. Evaluation Experiments for e-Markets. Zacharia and Maes [2000] proposed a set of experiments to evaluate their SPORAS model, a reputation model for e-markets. These experiments were then used and extended by other researchers to evaluate their own proposals and compare them with other reputation systems.

In Zacharia and Maes [2000] the SPORAS model was evaluated regarding the *speed or reputation convergence* (i.e., the time needed for reputation to reach its true value). It was also evaluated regarding how quickly it can adapt to the case of the traitor's attack: the case in which a trader behaves reliably for some time and then starts behaving dishonestly. The same experiments were used by Sabater and Sierra [2001] for their REGRET model and also by Carbo et al. [2003], who extended them with further scenarios (e.g., the case of collusion between sellers and buyers). The proposed set of experiments was used for the evaluation and comparison of reputation systems for e-commerce scenarios based on the utility gained by a single user.

2.2.2. Iterated Prisoner's Dilemma-Based Experiments. Game theory, and especially the PD problem, has been used by several research works for the evaluation of reputation systems. A PD game-based experiment for reputation system evaluation was first presented in the research work of Marsh [1994]. In this experiment, agents choose how to move in a cell grid using strategies and payoff matrices to aid their decisions. For a given interaction, agents choose their actions for a random situation. Payoffs are then made according to the given situation, and agents can adjust their trust values. Schillo et al. [2000] use an iterated PD game for presenting and evaluating their proposal for a reputation model. In that work, the proposed reputation system is based on a payoff matrix; partner selection for interaction through reputation evaluation; and publishing of interaction results, reputation information update, and payments. The evaluation of the proposed reputation system is done by simulating specific scenarios in which agents of selected agent profiles play the iterated game. The performance is measured through the *accumulated payoff* of the agents.

Mui et al. [2002] also propose a reputation system evaluation method through an iterated PD game played between specific types of agents that represent specific strategies. The authors evaluate different notions of reputation (*encounter-based, observed, group,* and *propagated reputation*) by simulating specific game scenarios among the various types of agents. The *evolution of the populations* of specific types of agents, each of which represents a specific strategy, is used to evaluate different notions of reputation. Lagesse et al. [2008] propose a utility-based reputation system for mobile P2P systems that takes into consideration energy consumption of mobile nodes. In their system, the nodes apply Nash Equilibrium strategies, and the selection of strategies is done according to the attacker model in consideration. The model is evaluated through simulation, in which a central peer recursively sends requests and calculates the experience received by other peers and the probabilities of other peers to attack.

Experiments evaluate the *average utility*, the *energy consumption,* and the *adaptability* of the model.

These PD-based evaluation proposals have been used to evaluate specific utility-based reputation systems and not to compare different reputation models.

2.3. Works Offering a Common Evaluation Framework

The various works aiming at evaluating and comparing reputation systems face the challenge of interoperability and choose either a theoretic comparison framework or a simulation-based comparison of a small number of reputation systems. The works found in the literature that propose a common framework for evaluation and comparison that will be applicable to various domains and reputation systems are thus categorized as theoretic and simulation-based approaches, as shown in Figure 1, and are described in the following sections. The distinction between the two categories is based on whether simulation experiments are included.

2.3.1. Theoretic Approaches. Theoretic methods for reputation systems' evaluation examine and/or compare reputation systems by either:

—considering their resilience to a set of specific threats and attempting a theoretic security analysis (e.g., Hoffman et al. [2009], Suryanarayana and Taylor [2006], and Gaur et al. [2010]).
—establishing a theoretic criteria framework for the evaluation of specific attributes or properties of reputation systems and using this framework for reputation system analysis and comparison (e.g., Ruohomaa et al. [2007], Noorian and Ulieru [2010], Lagesse [2012], and Vavilis et al. [2014]).

Theoretic approaches are thus based on either a threat analysis or a criteria framework and are presented here.

2.3.1.1. Threat Analysis-Based Theoretic Evaluation. The aim of this kind of evaluation is to identify the weak and strong points of reputation systems or give directions regarding which system is more resilient to specific threats. Hoffman et al. [2009] present a taxonomy of the attacks against reputation systems for P2P applications and an analysis framework that maps these attacks with specific components of reputation systems. They then study the resilience of a number of P2P reputation systems against these attacks based on their framework. Suryanarayana and Taylor [2006] also evaluate a number of decentralized reputation systems in a theoretic threat-analysis framework, whereas Gaur et al. [2010] present a similar analysis for the resilience of specific e-commerce reputation systems against some well-known reputation system attacks. Threat analysis-based theoretic evaluation of reputation systems gives useful insights and directions for reputation system designers; however, it does not provide a straightforward comparison of different reputation systems with measurable results in various scenarios.

2.3.1.2. Criteria Framework-Based Theoretic Evaluation. The objective of these works is to provide some criteria or measures that can be used for analysis. Because interoperability is a main obstacle for reputation systems comparison, some works propose a common conceptual interface for the implementation of reputation systems.

Ruohomaa et al. [2007] present an analysis and comparison of reputation systems regarding basic credibility criteria that refer to the *creation and content of recommendations*, the *selection of recommenders,* and the *reasoning for reputation estimation*. Noorian and Ulieru [2010] introduce a comparison framework that comprises several dimensions mapped to the major characteristics and requirements of reputation systems; specifically, to (i) *adaptability* (i.e., the possibility of adapting to dynamic open environments; e.g., to be scalable and to deal effectively with new entities); (ii) *context*

diversity checking (i.e., the possibility of integrating different context information); (iii) *accuracy* (i.e., the possibility of addressing some accuracy-related factors, such as time); and (iv) *reliability of reputation values* (i.e., the possibility of addressing various reliability-related issues). They use this framework for the performance analysis of a number of reputation systems with respect to the identified features. A similar approach was followed by Vavilis et al. [2014], who propose a framework for reputation system evaluation based on a number of reputation system requirements and the features needed to fulfil these requirements. They then use this framework for an analysis of the reputation metrics used in various reputation systems.

Hazard and Singh [2009] aim at enabling the evaluation and comparison of heterogeneous reputation systems by presenting a common conceptual interface for reputation systems that can be used for their implementation. They suggest four properties for reputation systems and related evaluation metrics that measure the *monotonic relationship* between actual behavior and reputation, the *accuracy* of reputation, the *convergence speed* of reputation values, and the *unambiguity* of a user's reputation in the long term. Finally, Lagesse [2012] suggests an evaluation framework for reputation systems that comprises evaluation metrics and a utility model and facilitates mathematical analysis of reputation mechanisms. The evaluation metrics are derived through modeling various kinds of peer behavior, peer connectivity, and the utility of benign and malicious peers. The authors describe generic models that can be extended to include various kinds of behavior, connectivity, and attack scenarios. The proposed metrics are *accuracy*, *convergence,* and *effectiveness* (a composite measure combining the previous two). The aim of this work is to facilitate the selection and development of suitable reputation mechanisms for different application and threat scenarios.

Criteria framework-based theoretic evaluation approaches provide a framework for comparison according to specific measures; however, they require an elaborative and critical analysis of the specific criteria set for the examined reputation systems and do not facilitate experiments in various scenarios and different environments.

2.3.2. Simulation-Based Approaches. Simulation-based approaches for defining a common evaluation and comparison framework have one or more of the following objectives:

—to simulate, evaluate, and compare reputation systems under similar scenarios and using specific measures (e.g., Schlosser et al. [2006], Kerr and Cohen [2009], and Hazard and Singh [2013]),
—to provide tools for reputation system implementation in prototype applications (e.g., Lagesse et al. [2009] and Suryanarayana et al. [2006]), or
—to provide tools/testbeds for reputation systems simulation, evaluation, and comparison (e.g., ART-Testbed [Fullam et al. 2006], TREET [Kerr and Cohen 2010; Irissappane 2012], To-SIM [Zhang et al. 2007], DART [Salehi-Abari and White 2012; Celestini et al. 2013a], TRMSim-WSN [Mármol and Pérez 2009], QTM [West et al. 2010], ATB [Jelenc et al. 2013]).

These works simulate a number of reputation systems as well as a number of specific attacks and examine reputation systems' vulnerabilities to these attacks. The advantage of simulation-based approaches is that they offer the possibility of experimentally comparing different reputation systems based on the same set of network settings and the same scenarios. Evaluation measurements can then be used for quantifiable comparisons. However, these approaches differentiate in many aspects related to the environment, the kind of attacks, the assumed organization of peers (centralized vs. decentralized), and the specific components of the reputation systems simulated.

2.3.2.1. Simulation-Based Comparison Frameworks. Schlosser et al. [2006] present a formal model for defining metrics for the representation of reputation systems and use it to model a number of global reputation metrics. Their model does not include other aspects, such as reputation query processing. Based on that model, a generic simulation framework for reputation metrics was implemented and used to compare different global reputation systems and find their strengths and weaknesses. The authors model specific agent types and simulate a number of threat scenarios. They thus compare these reputation systems using a number of basic criteria (*average reputation, reputation bias, successful transaction rate,* and *user profit*) to check to which degree the system can withstand attack. Similarly, Kerr and Cohen [2009] simulate a number of reputation systems for e-marketplaces, implement a number of attacks, and examine reputation systems' vulnerabilities to these attacks. In their subsequent work, Kerr and Cohen [2010] propose the reputation systems testbed TREET, which is presented in Section 2.3.2.3.

Hazard and Singh [2013] use a common conceptual interface and a set of four metrics, already presented in Hazard and Singh [2009], to simulate a number of reputation models in various threat scenarios. Their simulation framework is not described in detail. Liang and Shi [2008] also present a simulation framework for experimenting with a number of reputation systems factors that may be reputation model–related (e.g., the dissemination mode) or reputation-model independent (e.g., service quality offered by peers). They define several metrics for the performance evaluation of reputation estimation algorithms, and they present a comparison analysis of these algorithms, for evaluating the quality of a reputation model and for fine-tuning the various factors.

2.3.2.2. Tools for Reputation System Implementation in Prototype Applications. Lagesse et al. [2009] present DTT, a set of tools for reputation system implementation and evaluation in a variety of application environments. It consists of pluggable components that facilitate the tuning, reuse, extension, and combination of existing reputation mechanisms. However, Comparing different reputation systems can be done by using DTT to simulate a particular pervasive environment and implement distinct reputation mechanisms in this environment. In the presented simulations, the authors compare the use of different reputation mechanisms/components in a simulated pervasive environment regarding the *communication load*, *energy consumption,* and *confidence (accuracy)* obtained. Similarly to DTT, Suryanarayana et al. [2006] present PACE, which consists of a set of tools for the integration of trust models into a decentralized architecture. The authors show how PACE can be used to incorporate the implementations of different reputation systems within a prototype application. They simulate a number of threat scenarios and use their prototype applications for evaluation and comparison of the implemented reputation systems regarding their resilience to the examined threats.

2.3.2.3. Tools/Testbeds for Reputation Systems Simulation, Evaluation, and Comparison. Zhang et al. [2007] present TOSim, a simulator for distributed reputation systems built on the P2P network simulator PeerSim [PeerSim 2015]; their goal is to provide high extensibility and scalability, support for dynamic node behavior, and the simulation of various threat models. TOSim is based on the P2P file-sharing application scenario and on a number of highly trusted nodes.

The ART testbed [Fullam et al. 2006] was available as a competition platform in which researchers could compare their reputation models in the context of a painting appraisals community. It served also as a set of tools that allowed reputation system designers to perform customizable experiments. It offered user-specific metrics, such as the *accuracy* of a user's appraisals and the *consistency of this accuracy measure*, which affected a user's profit. It provided also system-specific statistics for the evaluation of

the social benefit offered by a reputation system. However, it imposed limitations on the simulated reputation systems and specific assumptions that were not applicable for all reputation systems. Support for the ART testbed is no longer available.

More recently, Irissappane et al. [2012] propose a testbed for reputation systems that aims at offering a comprehensive evaluation solution for reputation systems by considering their approaches toward unfair rating detection. Their testbed supports simulation of various environments (real environments or simulated ones) and various attack models. It thus allows the comparison of reputation systems under the same attack model but in different environments, or under different attack models but in the same environment. The evaluation of reputation systems is done using various *robustness* metrics, such as the *transaction volume difference between honest and dishonest sellers* and the *number of unfair ratings needed to change a target's reputation*.

Kerr and Cohen [2010] propose the reputation systems testbed TREET, which models a general e-marketplace scenario and can be used for reputation systems evaluation. It allows the implementation of different reputation systems, agent profiles, and test scenarios and also allows testing of particular components of reputation systems. The authors aim at offering the possibility for objective benchmarking and comparison of various implementations. Some of the limitations of TREET are that it is designed specifically for marketplaces and that evaluation considers only the *utility of a single type of entity* and not the performance or accuracy of the reputation system as a whole. The evaluation metric used is the ratio of *sales/profits between honest* and *cheating agents*.

Celestini et al. [2013a, 2013b] provide a software framework for the evaluation of reputation systems that are based on the notion of probabilistic trust. Specifically, the authors describe a tool for prototyping and evaluation of reputation systems that explicitly takes into account the networked execution environment. This tool allows the implementation of reputation models, the setup of specific networks, and the setup of specific scenarios. It can thus be used to compare the performance of a reputation system in various configurations, as well as to compare different reputation systems. In their work, they consider a centralized architecture for the collection of ratings and for the search of resource providers. They consider experimenting with decentralized architectures in their future work. A further limitation of this tool is that it currently targets only probability theory–based reputation systems.

SIFT [Suryanarayana and Taylor 2007] is a simulation framework for exploring and fine-tuning reputation systems. In their simulations, the authors experiment with various parameters related to (i) reputation formation, (ii) application context, and (iii) simulation, aiming at evaluating the strengths and weaknesses of different settings. The authors use the simulation results to discuss the impact of different trust settings on several reputation models proposed in the literature. They aim at helping a reputation system designer to make the right choices regarding the selection and refinement of various trust parameters when using a specific reputation model.

DART [Salehi-Abari and White 2012] is a framework for reputation system analysis based on PD games, which also provides a set of tools and evaluation metrics for the simulation and evaluation of reputation systems. Transactions are assigned as either "cooperation" or "defection," which can have different meanings according to the domain. DART enables the implementation of different reputation models and their evaluation against various attacks. It is thus domain independent and flexible. The main evaluation metrics are the *utility* of a user and the *number of unsuccessful connections* made by users. A drawback is that the evaluation depends not only on a specific trust model but also on various decision mechanisms that have to be made regarding recommender's selection and various other policies. Furthermore, implemented systems and experiments in DART are not publicly available to enable further evaluations and comparisons.

In addition to the aforementioned works, there are also some simulation-based CEFs that focus on offering the possibility to other researchers of easily implementing a reputation system and comparing it with other implemented reputation systems. Such works include TRMSim-WSN [Mármol and Pérez 2009], QTM [West et al. 2010; QTM 2014], and ATB [Jelenc et al. 2013]; these make reputation system comparison easier and more transparent.

Mármol and Pérez [2009] present TRMSim-WSN, a simulation application that offers a general API for the implementation of new reputation models, and they base their evaluation on a number of evaluation criteria. It also allows users to adjust several parameters for the simulation of specific threat scenarios (e.g., oscillating behavior and collusion). In this simulator, several known reputation systems have been implemented, and they are available for comparison with new implementations of reputation systems. The evaluation metrics that TRMSim-WS uses are (i) the *average length* of all the paths found by every user of every simulated network, (ii) the *accuracy of the model*, and (iii) the *energy consumed* by the reputation systems in the case of wireless sensor networks. The reputation systems can be evaluated using these metrics under a number of threat scenarios, as presented in Mármol and Pérez [2011].

QTM [West et al. 2010; QTM 2014] proposes the use of a single metric for the evaluation of different simulated reputation systems and uses it to evaluate a reputation algorithm under different threat scenarios. This metric is the *hit rate* of honest users (e.g., the percentage of their successful transactions), and the assumed environment is a file-sharing community. Furthermore, the *convergence rate* of the simulated reputation algorithms can also be evaluated. The default testbed of QTM contains six user models for the simulation of a variety of attacks, and the authors state that more user models can be easily implemented to simulate more complex malicious strategies. QTM has been used by An et al. [2013] to evaluate their proposed reputation system.

Jelenc et al. [2013] propose ATB, a testbed for the evaluation of reputation systems that is built on the open-source and widely used simulation platform Repast [North et al. 2007]; it emphasizes the significance of the decision-making mechanism in the evaluation of reputation systems through simulation. Reputation systems can be implemented as plugins, which integrate a trust model with scenarios and evaluation metrics. Various reputation systems have been implemented, and the evaluation results mostly concern the comparison of reputation systems with the use of different decision-making mechanisms for the selection of counterparties in transactions.

From our survey, we can see that most work toward providing a CEF for reputation systems benchmarking are not finalized; instead, the researchers aim at enhancing their CEFs in their future work to provide easier and more reliable reputation system evaluation and comparison.

3. TOWARD A GENERALLY AGREED-ON COMMON FRAMEWORK FOR EVALUATION

In this section, we present the limitations of current works on simulation-based CEFs and define the desirable characteristics of a commonly agreed-on CEF. We also analyze existing works according to their conformance to these characteristics and present some factors for their enhancement.

3.1. Limitations of Current Works on Simulation-based CEFs

Based on our survey on the various simulation-based CEF proposals, we conclude that the factors that differentiate them and limit their use as general frameworks for evaluation and comparison are the following:

—*The supported simulation environment characteristics*: For example, the population size, simulation of the dynamic behavior of participants, simulation of social relationships, etc.

—*The evaluation metrics*: Works on CEFs use different evaluation metrics that concern either a specific user (user metrics) or the system as a whole (system metrics). Examples of the evaluation metrics used in the various works are *convergence* (Liang and Shi [2008], Celestini et al. [2013a, 2013b], and Hazard and Singh [2013]), *user profit* (Kerr and Cohen [2009]), *accuracy* (Mármol and Pérez [2009]), *transaction volume difference between honest and dishonest service providers* (Kerr and Cohen [2010]), and *hit rate* (West et al. [2010]). A consensus on the definition and estimation of metrics is lacking.
—*The addressed attack models*: Usually, these attack models are simulated with the design of suitable user profiles and the use of suitable scenarios. Some simulation approaches (e.g., Schlosser et al. [2006] and West et al. [2010]) use sophisticated user profiles or/and threat scenarios; for example, in QTM [West et al. 2010; QTM 2014] a large number of user profiles are simulated (good, purely malicious, single-dimension malicious, disguised malicious, Sybil attacker, and collective malicious) and are used to simulate various attack scenarios. On the other hand, Celestini et al. [2013a, 2013b] study the performance of reputation systems in various configurations without explicitly describing any threat scenarios.
—*The kind of reputation systems / reputation system components being addressed*: Some CEF approaches address specific kinds of reputation systems; for example, only centralized (Celestini et al. [2013a, 2013b]) or decentralized (Zhang et al. [2007]) or reputation systems specialized for e-commerce applications (Kerr and Cohen [2010]). Other CEFs consider only specific components, such as the reputation metric, as in Schlosser et al. [2006].
—Last, *some evaluations depend on the decision mechanisms used in a reputation system* [Jelenc et al. 2013], and these mechanisms are not always explicitly modeled in CEFs: Consequently, evaluation under different decision mechanisms requires different reputation system implementations. Furthermore, comparing reputation systems is not straightforward; it may be biased due to different policies.

These variances hinder the establishment of a generally agreed-on CEF that would enable the objective evaluation of most reputation systems and a reliable comparison between them.

3.2. Desirable Characteristics of a CEF

In this section, we formulate a set of characteristics/properties that are desirable for a generally accepted CEF. We use these characteristics to analyze the existing simulation-based CEF proposals described in Section 2.2.1. We then elaborate on how these desirable characteristics can be achieved by presenting a list of factors that could be taken into account in the implementation of a CEF, and we discuss possible conflicts and tradeoffs between the proposed characteristics.

(1) **Standardization**: A CEF needs to incorporate an *abstraction* of reputation models that can be used for reputation systems implementation (e.g., through a common interface [Hazard and Singh 2013], [Mármol and Pérez 2009]) and a number of well-defined evaluation metrics that will be applicable to most reputation systems.
(2) **Independence of the reputation system characteristics**: A CEF should be independent of market and transaction characteristics; at the same time, it should be able to support various market and service models. It should also enable simulation of reputation systems independently of their architecture (centralized or decentralized) or the kind of reputation metric used (e.g., probabilistic, fuzzy logic-based, or deterministic).

(3) **Flexibility**: A CEF should be flexible regarding its configuration and its evaluation possibilities, and thus be able to adapt to a reputation system's unique characteristics. More specifically, it should enable the simulation of:
—a wide range of attack models, from simple to more sophisticated forms of malicious or strategic user profiles that interact with each other;
—different aspects of the reputation system (e.g., the reputation estimation metric, the social relations when needed, etc.);
—the target environment, by adapting that environment's characteristics (e.g., population size, context of services, dynamics of the participants, transaction scenarios, etc.).

It should also enable the evaluation of a reputation mechanism under different decision mechanisms regarding the selection of counterparties, as pointed out in Jelenc et al. [2013].

Flexibility could be achieved through a modular simulation environment in which different modules (which may implement protocols, application scenarios, user types, etc.) could be tested separately and could be easily replaced through configuration files, as proposed in Zhang et al. [2007].

(4) **Ease of implementation of new reputation systems and new tests**: Researchers should be able to easily implement their own tests for evaluating particular aspects of their reputation systems. Suitable abstractions for reputation systems, as well as implementation tools and documentation, could facilitate this task.

(5) **Availability of existing implementations of reputation system tests in order to produce reliable comparisons**: Existing implementations of reputation systems and tests could be useful for other researchers, too, so it would be helpful if they were made available for reuse. Reusability of tests can also enable the reproducibility of evaluation results, which, in turn, can facilitate reputation system comparison and fine-tuning.

A generally accepted CEF should address these characteristics at a level that offers standard and objective evaluation, extensibility, ease of use, and availability.

3.2.1. Analysis of Existing CEFs Regarding the Proposed Characteristics. In Table II, we present the conformance level of the simulation-based CEF proposals described in Section 2.3.2 as they relate to the desirable characteristics just listed and according to information provided by the authors. In the same table, we also include the suggested evaluation metrics in each of the studied CEF proposals as an indicator of the evaluation approach followed.

More specifically, the conformance of each CEF work to the desirable characteristics is judged by answering the following questions:

Standardization: Does the examined CEF provide an abstraction for reputation systems? Are there well-defined evaluation metrics? Are there well-defined attack scenarios?

Independence of the reputation system characteristics: Does the examined CEF have limitations regarding the architecture, the type of evaluation metrics, or the application context that it supports?

Flexibility: Is configuration and/or modularization enabled regarding the simulation parameters, the threat model, and the various components of reputation systems?

Ease of implementation of new reputation systems and new tests: Are there guidelines, source code, and documentation for the implementation of the needed simulations?

Table II. Analysis of CEFs Regarding Evaluation Metrics and Conformance to Desired Characteristics

		Proposed evaluation metrics (user/system)	Standardization	Independence of the reputation system characteristics	Flexibility	Ease of implementation of new reputation systems and new tests	Availability of existing implementations of reputation system tests
Simulation-based comparison frameworks	Schlosser et al. [2006]	*User type metrics*: reputation bias, hit rate, profit per transaction	Partially	Partially; it concerns only global metrics	No	✓	Currently not available
	Macau	*System metrics*: monotonicity, accuracy, convergence, unambiguity	✓	✓	No information provided	✓	✓
	Liang and Shi [2008]	*System metrics*: convergence time, selection hit, workload, correction rate, storage cost, running time, benefit speedup, efficient factor	Partially, because it is built on a widely used simulator	✓	✓	Partially	Currently not available
Tools/ testbeds for reputation systems simulation, evaluation and comparison (continued)	Zhang et al. [2007]	No specific metrics suggested	Partially, because it is built on a widely used P2P simulator	Only for P2P overlay reputation systems	✓	✓	Currently not available
	ART testbed	*User metrics*: average accuracy, consistency of accuracy *System metrics*: statistics for earnings, transactions, messages,	Partially	✓	Partially	Partially	Currently not available

(Continued)

Table II. Continued

		Proposed evaluation metrics (user/system)	Standardization	Independence of the reputation system characteristics	Flexibility	Ease of implementation of new reputation systems and new tests	Availability of existing implementations of reputation system tests
	SIFT	No specific metrics suggested	No	✓	✓	Partially	Currently not available
	Kerr and Cohen [2010]	*User metric*: Profit from sales	Partially; it refers to e-markets	✓	✓	✓	✓
	Irissappane et al. [2012]	*System metrics*: mean absolute error, precision and recall, transaction volume difference between honest and dishonest sellers, number of unfair ratings required to change a target's reputation	✓ As a future step	✓	✓	✓ Through sophisticated interfaces (future work)	Future plan
	DART	*User metrics:* utility based on payoff matrix, number of dropped connections*System tool*: graph visualization of interactions	✓ Domain independent	✓	✓	Partially (no specific instructions are given)	Currently not available

(Continued)

Table II. Continued

		Proposed evaluation metrics (user/system)	Standardization	Independence of the reputation system characteristics	Flexibility	Ease of implementation of new reputation systems and new tests	Availability of existing implementations of reputation system tests
	Celestini et al. [2013a, 2013b]	*System metric*: convergence in terms of number of rating needed	No	No; currently deployed only in centralized architecture and for probabilistic-reputation models	Flexibility regarding configuration	Partially	Two reputation models have been implemented and have become available
	TRMSim-WS	*System metrics*: accuracy, path length	✓	✓	✓	✓	✓
	QTM	*System metric*: hit rate	✓	✓	✓ According to authors, new attacks may be simulated	✓	✓
	ATB	*User metrics*: accuracy, utility, opinion cost	✓ Built on Repast framework [North et al. 2007]	✓	✓	✓	✓ Source code and instructions available
Tools and architectures for rep. sys. development in prototype applications	DTT	*System metrics*: # messages, confidence level	No information provided	✓	✓	✓ Through specific components	Currently not available
	PACE	No specific metrics suggested	Partially	✓	✓ Supports reuse	Partially; it requires a prototype application	Currently not available

Availability of existing implementations of reputation system tests: Do the authors provide the code for existing implementations of simulated reputation systems or reputation system components so that they can be reused?

Based on our analysis, we can draw a first inference that the works by Mármol and Pérez [2009], West et al. [2010], and Jelenc et al. [2013] are viable candidates that researchers should consider extending to establish generally accepted CEFs; these works offer the code for their simulators, available reputation system simulations, and instructions on implementing the simulation of a new reputation system.

We note that, in order for these or other CEFs to be considered optimal, further work is needed. More specifically, the listed desirable characteristics could be further broken down to enable a more detailed analysis of CEF works (suggested subcriteria for further analysis are presented in the next section). Further criteria could be used, including those related to application-specific requirements (adhering thus to the "flexibility" characteristic), such as the possibility of evaluating privacy levels in reputation systems for ubiquitous applications. Such an extended analysis, which is beyond the scope of this survey, can lead to stronger conclusions about the selection of a CEF that will fulfill particular requirements (e.g., a CEF for the evaluation of reputation systems for e-commerce communities) or that will be considered as a generally accepted CEF. Finally, the verification of the analysis results will require feedback from the research community about the CEFs' practical usage.

3.2.2. Factors to Be Taken into Account for the Definition and Implementation of a CEF. We elaborate on the discussed properties by suggesting a number of factors that should be taken into consideration when defining and implementing a CEF. We note that the factors presented here constitute neither an exhaustive nor exclusive list, but are instead some indicative suggestions on how to achieve the desired properties of a generally accepted CEF.

Abstract Model for Reputation Systems. A CEF should provide an abstraction for modeling reputation systems that could facilitate the simulation of different reputation systems, taking into account various aspects of their functionality.

Standard Evaluation Metrics and Possibility of Using Additional Evaluation Metrics. A CEF should provide a set of standard evaluation metrics that will be commonly accepted and also additional evaluation metrics that will apply in particular environments or reputation systems. The possibility of implementing new metrics for the evaluation of new aspects of particular reputation systems should be provided. The evaluation metrics should concern either the individual entities (user metrics) or the whole community (system metrics), according to the application context and to the cooperativeness level of the community. For example, in e-commerce applications, user metrics may be needed. For a competitive environment where users compete with each other for the same resources/services, accuracy for the whole system should be measured in scenarios that consider unfair ratings and collusions.

Modeling Utility and Cost of Recommendations and Services. Recommendation economics modeling can allow the integration of both the overhead incurred and the utility obtained by a reputation mechanism in the reputation system's simulation, thus resulting in a more accurate evaluation [Lagesse 2012]; for example, the tradeoff between accuracy obtained from a large number of recommendations and the relative cost should be considered. Additionally, modeling utility and the cost of service transactions (e.g., utility and cost of an e-commerce transaction) could allow more accurate simulation and evaluation of a reputation system, especially in the case of user-centric evaluation [Kerr and Cohen 2010].

Modeling Social Relationships. When the community is characterized by social relations among its members that affect the reputation system, these social relationships should be modeled both for the appropriate simulation of such a reputation system and the more accurate evaluation of its resilience against collusion attacks.

Modeling Dynamic Behavior of Users. Dynamic behavior of entities should be modeled for the simulation of (i) communities where entities can enter and leave the system freely, either permanently or temporarily, and (ii) the mobility of nodes in mobile ad-hoc networks in which the topology changes.

Various Threat Scenarios. Various threat scenarios should be available for tests. This may be done through modeling various user profiles regarding their transactional or recommending behaviors. Different types of user profiles should be available, and it should be easy to define new user profiles. These user profiles can be used to set up relative threat scenarios through the configuration of the percentages of particular types of malicious users. Apart from user profiles, alternative ways of testing attacks can be used, such as controlled experiments with specific users and a combination of metrics, as in Macau [Hazard and Singh 2013], or by specifying percentages of wrong information and of false information per unit of time [Sabater 2004].

Time Modeling. The concept of time should be explicitly modeled because it can be used to define time periods for simulated scenarios and to weight reputation information according to its recency.

Interaction Modeling. For enabling the simulation of a realistic network of entities that interact with each other in a specific application context, the parameters available for configuration include the number of participating entities, the number of interactions, and the interaction frequency. These factors should be considered when deciding how to weight direct reputation information in relation to indirect information when simulating and evaluating a reputation model. For example, a combination of a low population of entities with a high number and high frequency of interactions results in a large number of direct experiences. On the other hand, a combination of a large number of participants with a low number of interactions and low interaction frequency needs more indirect information that should be weighted higher when estimating a user's reputation. When evaluating a reputation system in the former case, more emphasis should be given to scenarios with oscillating transactional user behavior, whereas, when evaluating the latter case, emphasis should be placed on resistance to unfair or wrong recommendations. These parameters can also be used to simulate scenarios of network problems that hinder interactions.

Table III presents the described factors together with the related properties of CEFs and representative CEF proposals that use them.

3.2.3. Discussion on the Proposed Characteristics of a CEF. The proposed desirable characteristics of a CEF aim at enabling the objective evaluation, fine-tuning, and comparison of reputation systems, as well as the easy implementation and extension of reputation system simulations. We presented them qualitatively in the form of requirements and suggested factors, and we analyzed existing CEFs regarding the ways in which they deal with these characteristics. The study of the proposed desirable characteristics for CEFs and the analysis of existing CEFs reveal a number of challenges and particularities that we address in this section.

For example, there might seem to be a conflict between *standardization* and *flexibility*. On the one hand, *standardization* necessitates the ability to model a reputation model according to an abstract model and use already defined evaluation metrics and scenarios. On the other hand, *flexibility* requires the ability to implement different

Table III. Factors for Enhancing Desirable Characteristics of CEFs

Factor	Desirable Property	Representative CEFs
Abstract model for reputation system	Standardization, ease of implementation of new reputation systems and new tests	Schlosser et al. [2006], Liang and Shi [2008], Mármol and Pérez [2009], Lagesse [2012]
Standard evaluation metrics and possibility of using additional evaluation metrics	Standardization, ease of implementation of new reputation systems and new tests, flexibility	Mármol and Pérez [2009], West et al. [2010], Jelenc et al. [2013]
Modeling utility and cost of recommendations and services	Flexibility	Liang and Shi [2008], Jelenc et al. [2013], Lagesse [2012], Kerr and Cohen [2010]
Modeling social relationships	Flexibility	Salehi-Abari and White [2012], Liang and Shi [2008], Jelenc et al. [2013]
Modeling dynamic behavior of users	Flexibility	Salehi-Abari and White [2012], TRM-Sim, Liang and Shi [2008]
Various threat scenarios	Flexibility	West et al. [2010], Mármol and Pérez [2009], Jelenc et al. [2013]
Time modeling	Flexibility	Jelenc et al. [2013], Liang and Shi [2008]
Interaction modeling	Flexibility	West et al. [2010]

scenarios and to evaluate a reputation system using the most suitable evaluation metrics for a particular scenario. However, suitable tradeoffs should made so that both characteristics are satisfied. We should thus be able to both use a set of standard simulation scenarios and also to create new scenarios according to the special nature of specific reputation systems that should be taken into account in their evaluation and comparison. For example, reputation systems that estimate global reputation values after every transaction and thus require high computational or communication cost are not expected to scale well to large populations with frequent transactions. Therefore, they should be evaluated using suitable simulation scenarios with moderate transaction frequency.

Flexibility should also be complementary to *independence of reputation system characteristics*; reputation systems should be able to be simulated independently of their specificities, while, at the same time, these specificities should be able to be included if needed. For example, an ideal CEF should be able to evaluate reputation systems independently of their architecture (centralized or decentralized). At the same time, optimally, it should enable the evaluation of specific reputation system performance aspects (such as scalability), taking into account related characteristics such as architecture.

Availability of existing implementations of reputation system simulations is a very useful and vital characteristic of a CEF because it allows the easy comparison of reputation systems. It also permits a critical review of available simulations, which is required for an objective and meaningful comparison. This critical analysis is also required when deciding to reuse available components for the implementation of new tests.

We note that, even if an ideal CEF is available, the reliable evaluation of a reputation system requires decisions regarding how to suitably model the reputation system and transactional context and decisions about the most relative scenarios that will be examined. For example, for the evaluation of a reputation system that requires more accuracy (and thus incurs higher communication costs), the tradeoff between accuracy and scalability should be addressed with the appropriate simulation scenarios and

tests. The desirable characteristics of a CEF, when present, facilitate the easy and reliable implementation of the necessary decisions and lead to objective evaluation, comparison, and fine-tuning of reputation systems.

Finally, we would like to note that recent works toward constructing reputation models in a standardized way could be investigated in the context of defining a widely accepted CEF. Such works include Hillebrand and Coetzee [2013], which describes the requirements for implementing a reputation system as a service, and Sänger et al. [2015], which proposes a component repository for reputation system components at both conceptual and implementation levels. The potential of these works to provide reputation system implementation standardization could be used in conjunction with works on CEFs. More specifically, a CEF could integrate the possibility for standardized implementations, thus enhancing the specified desirable characteristics.

4. CONCLUSION

Motivated by the need for reliable and objective evaluation of reputation systems in various application environments, we exhaustively examined the evaluation mechanisms available in the literature. Based on the evaluation requirements and shortcomings of the current approaches, the aim of this analysis is to provide directions to researchers for choosing a suitable evaluation mechanism and simulation tool, on one hand, and for establishing a commonly agreed-on evaluation framework on the other. We thus classified current works on reputation system evaluation mechanisms in a basic taxonomy, according to the spectrum of usage: custom-made evaluation mechanisms for individual reputation systems, common experiments under restricted scenarios, and evaluation works that aim at constituting commonly agreed-on evaluation frameworks. We identified various evaluation approaches in individual reputation system proposals and focused on works on defining a testbed of reputation systems that would enable easy simulation of reputation systems and their objective evaluation and comparison.

We found that researchers, although having identified the need for a commonly agreed evaluation and comparison framework, have not yet reached such an agreement and that the various works are presented as proposals that need more feedback from other researchers and future work for enhancement and completion. In order to contribute to such a CEF definition, we suggest the desired properties of an envisioned CEF and analyze current proposals for CEFs regarding these properties. We also discuss how these properties can be achieved and their relative challenges.

To the best of our knowledge, our proposal for a framework of desirable CEF characteristics that can be used to analyze existing CEF proposals and define a generally accepted CEF is a novel work in this direction. Therefore, the contribution of the article is threefold:

(1) It provides insights for designers in choosing a suitable evaluation method for their reputation systems—based either on their own criteria or experiments or on an available CEF—through the presented taxonomy and analysis of the evaluation methods of existing reputation systems.
(2) It helps reputation system designers seeking a CEF that will provide them with tools for objective evaluation and/or fine-tuning to choose a suitable CEF among those available. This can be done through our presentation and analysis of existing simulation-based CEFs.
(3) It provides directions for the design of a generally agreed-on CEF through a framework of suggested properties, a discussion about how they can be achieved, and an analysis of existing CEF proposals regarding the level of their conformance to these properties.

Although critical thinking is always necessary in deciding on the reliable evaluation of reputation systems, we believe that our work facilitates these decisions and can thus be used as a roadmap for the design and implementation of suitable evaluation experiments and for the design of a generally accepted CEF. Reliable evaluation can lead, in turn, to fine-tuning, enhancement, and improvement of reputation systems and also to the ability to compare different reputation systems and choose the most suitable one for a given application environment.

Our future plans include further work on benchmarking of reputation systems: on defining and implementing a CEF that will be grounded on or extend current CEF approaches and will incorporate the aforementioned desirable characteristics. Subsequently, we plan to use such a CEF to thoroughly experiment with the evaluation of the reputation algorithms we have developed [Koutrouli and Tsalgatidou 2013] and other reputation systems in various application environments, thus contributing to the design of optimal reputation systems for specific contexts.

REFERENCES

Karl Aberer and Zoran Despotovic. 2001. Managing trust in a peer-2-peer information system. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01)*, Henrique Paques, Ling Liu, and David Grossman (Eds.). ACM, New York, NY, 310–317. DOI: http://doi.acm.org/10.1145/502585.502638

Florina Almenárez, Andrés Marín, Celeste Campo, and Carlos García-Rupo. 2004. PTM: A pervasive trust management model for dynamic open environments, privacy and trust. In *Proceedings of the 1st Workshop on Pervasive Security and Trust*, Boston, MA, Aug. 2004.

Do-sik An, Byong-lae Ha, and Gi-hwan Cho. 2013. A robust trust management scheme against the malicious nodes in distributed p2p network. *International Journal of Security and Its Applications* 7, 3, 317–326, May 2013.

Sonja Buchegger and Jean-Yves Le Boudec. 2004. A robust reputation system for P2P and mobile adhoc networks. In *Proceedings of the 2nd Workshop on the Economics of Peer-to-Peer Systems*. Cambridge MA, June 2004.

Javier Carbo, Jose M. Molina, and Jorge Davila. 2003. Trust management through fuzzy reputation. *International Journal of Cooperative Information Systems* 12, 1 (2003).

Alessandro Celestini, Rocco De Nicola, and Francesco Tiezzi. 2013a. Network-aware evaluation environment for reputation systems. In *Proceedings of the 7th IFIP WG 11.11 International Conference (IFIPTM'13)*. 231–238.

Alessandro Celestini, Rocco De Nicola, and Francesco Tiezzi. 2013b. Network-aware evaluation environment for reputation systems (long version), IMT Institute for Advanced Studies Lucca, Lucca, Italy, CSA Tech. Rep. #05/2013.

eBay. 2014. Homepage. Retrieved from http://www.ebay.com.

Karen K. Fullam, Tomas Klos, Guillaume Muller, Jordi Sabater-Mir, K. Suzanne Barber, and Laurent Vercouter. 2006. The agent reputation and trust (ART) testbed. In *Proceedings of the 4th International Conference on Trust Management (iTrust'06),* Ketil Stølen, William H. Winsborough, Fabio Martinelli, and Fabio Massacci (Eds.). Springer-Verlag, Berlin, 439–442. DOI: http://dx.doi.org/10.1007/11755593_32

Vibha Gaur, Neeraj Kumar Sharma, and Punam Bedi. 2010. Evaluating reputation systems for agent mediated e-commerce. In *Proceedings of ACEEE Conference: International Conference on Advances in Computer Science (ACS'10)*.

Christopher J. Hazard and Munindar P. Singh. 2009. Reputation dynamics and convergence: A basis for evaluating reputation systems, North Carolina State University, Department of Computer Science, Raleigh, NC, Tech. Rep. TR-2009-19, Nov. 2009.

Christopher J. Hazard and Munindar P. Singh. 2013. Macau: A basis for evaluating reputation systems. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13)*, Francesca Rossi (Ed.). AAAI Press, 191–197.

Channel Hillebrand and Marijke Coetzee. 2013. Towards Reputation-as-a-Service. *Information Security for South Africa 2013*, 1, 8 (Aug. 2013), 14–16. DOI: http://dx.doi.org/10.1109/ISSA.2013.6641047

Kevin Hoffman, David Zage, and Cristina Nita-Rotaru. 2009. A survey of attack and defense techniques for reputation systems. *ACM Computer Survey* 42, 1, Article 1 (December 2009), 31 pages. DOI: http://dx.doi.org/10.1145/1592451.1592452

Chenlin Huang, Huaping Hu, and Zhiying Wang. 2006. A dynamic trust model based on feedback control mechanism for P2P applications. In *Proceedings of the 3rd International Conference on Autonomic and Trusted Computing (ATC'06)*, Laurence T. Yang, Hai Jin, Jianhua Ma, and Theo Ungerer (Eds.). Springer-Verlag, Berlin, 312–321. DOI:http://dx.doi.org/10.1007/11839569_30

Kuan Lun Huang, Salil S. Kanhere, and Wen Hu. 2012. A privacy-preserving reputation system for participatory sensing. In *Proceedings of the 2012 IEEE 37th Conference on Local Computer Networks (LCN'12)*. IEEE Computer Society, Washington, DC, 10–18. DOI:http://dx.doi.org/10.1109/LCN.2012.6423585

Athirai Aravazhi Irissappane, Siwei Jiang, and Jie Zhang. 2012. Towards a comprehensive testbed to evaluate the robustness of reputation systems against unfair rating attacks. In *Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP'12)*.

David Jelenc, Ramón Hermoso, Jordi Sabater-Mir, and Denis Trček. 2013. Decision making matters: A better way to evaluate trust models. *Knowledge-Based Systems* 52 (November 2013), 147–164. DOI:http://dx.doi.org/10.1016/j.knosys.2013.07.016

Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. 2003. The Eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the 12th International Conference on World Wide Web (WWW'03)*. ACM, New York, NY, 640–651. DOI:http://doi.acm.org/10.1145/775152.775242

Reid Kerr and Robin Cohen. 2009. Smart cheaters do prosper: Defeating trust and reputation systems. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2 (AAMAS'09)*, Vol. 2. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 993–1000.

Reid Kerr and Robin Cohen. 2010. TREET: The trust and reputation experimentation and evaluation testbed. *Electronic Commerce Research* 10, 3 (Aug. 2010), 271–290. DOI:http://dx.doi.org/10.1007/s10660-010-9056-y

Eleni Koutrouli and Aphrodite Tsalgatidou. 2012. Taxonomy of attacks and defense mechanisms in P2P reputation systems — Lessons for reputation system designers. *Computer Science Review* 6, 2–3 (May 2012), 47–70.

Eleni Koutrouli and Aphrodite Tsalgatidou. 2013. Credible recommendation exchange mechanism for P2P reputation systems. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC'13), Trust, Reputation, Evidence and Other Collaboration Know-how (TRECK) Track*. ACM, New York, NY, 1943–1948. DOI:http://doi.acm.org/10.1145/2480362.2480724

Brent Lagesse. 2012. Analytical evaluation of P2P reputation systems. *International Journal of Communication Networks and Distributed Systems* 9, 1/2 (July 2012), 82–96. DOI:http://dx.doi.org/10.1504/IJCNDS.2012.047897

Brent Lagesse, Mohan Kumar, and Matthew Wright. 2008. AREX: An adaptive system for secure resource access in mobile p2p systems. In *Proceedings of the 2008 8th International Conference on Peer-to-Peer Computing (P2P'08)*. IEEE Computer Society, Washington, DC, 43–52. DOI:http://dx.doi.org/10.1109/P2P.2008.46

Brent Lagesse, Mohan Kumar, Justin Mazzola Paluska, and Matthew Wright. 2009. DTT: A distributed trust toolkit for pervasive systems. In *Proceedings of the 2009 IEEE International Conference on Pervasive Computing and Communications (PERCOM'09)*. IEEE Computer Society, Washington, DC, 1–8. DOI:http://dx.doi.org/10.1109/PERCOM.2009.4912754

Zhengqiang Liang and Weisong Shi. 2005. PET: A personalized trust model with reputation and risk evaluation for P2P resource sharing. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, Volume 07. IEEE Computer Society, Washington, DC, 201. DOI:http://dx.doi.org/10.1109/HICSS.2005.493

Zhengqiang Liang and Weisong Shi. 2008. Analysis of ratings on trust inference in open environments. *Performance Evaluation* 65, 2 (February 2008), 99–128.

Félix Gómez Mármol and Gregorio Martínez Pérez. 2009. TRMSim-WSN, trust and reputation models simulator for wireless sensor networks. In *Proceedings of the 2009 IEEE International Conference on Communications (ICC'09)*. IEEE Press, Piscataway, NJ, 915–919.

Félix Gómez Mármol and Gregorio Martínez Pérez. 2010. Towards pre-standardization of trust and reputation models for distributed and heterogeneous systems. *Computer Standards & Interfaces*, 32, 4 (June 2010), 185–196. DOI:http://dx.doi.org/10.1016/j.csi.2010.01.003

Felix Gomez Mármol and Gregorio Martinez Pérez. 2011. Trust and reputation models comparison. *Internet Research* 21, 2 (2011), 138–153. DOI:http://dx.doi.org/10.1108/10662241111123739

Félix Gómez Mármol and Gregorio Martínez Pérez. 2012. TRIP, a trust and reputation infrastructure-based proposal for vehicular ad hoc networks. *Journal of Network and Computer Applications* 35, 3 (May 2012), 934–941. DOI: http://dx.doi.org/10.1016/j.jnca.2011.03.028

Sergio Marti and Hector Garcia-Molina. 2004. Limited reputation sharing in P2P systems. In *Proceedings of the 5th ACM Conference on Electronic Commerce (EC'04)*. ACM, New York, NY, 91–101. DOI:http://doi.acm.org/10.1145/988772.988787

Sergio Marti and Hector Garcia-Molina. 2006. Taxonomy of trust: Categorizing P2P reputation systems. *Computer Networks* 50, 4 (March 2006), 472–484. DOI:http://dx.doi.org/10.1016/j.comnet.2005.07.011

Stephen Paul Marsh. 1994. Formalising trust as a computational concept. PhD Thesis, Dept. of Computing Science and Mathematics, University of Stirling, Stirling, UK, 1994.

Lik Mui, Ari Halberstadt, and Mojdeh Mohtashemi. 2002. Evaluating reputation in multi-agents systems. In *Proceedings of the 2002 International Conference on Trust, Reputation, and Security: Theories and Practice (AAMAS'02)*, Rino Falcone, Suzanne Barber, Larry Korba, and Munindar Singh (Eds.). Springer-Verlag, Berlin, 123–137.

Zeinab Noorian and Mihaela Ulieru. 2010. The state of the art in trust and reputation systems: A framework for comparison. *Journal of Theoretical and Applied Electronic Commerce Research* 5, 2 (Aug. 2010), 97–117.

Michael J. North, Tom R. Howe, Nicholson T. Collier, and Richie Vos. 2007. A declarative model assembly infrastructure for verification and validation. In *Proceedings of the Advancing Social Simulation: The First World Congress*, 129–140, 2007.

PeerSim. 2015. Homepage. Retrieved from http://peersim.sourceforge.net/#pubs.

QTM. 2014. QTM: Trust Simulator. Retrieved from http://rtg.cis.upenn.edu/qtm/p2psim.php3, last accessed on 28/9/2015.

Tanya Roosta, Marci Meingast, and S. Shankar Sastry. 2006. Distributed reputation system for tracking applications in sensor networks. In *Proceedings of the Annual International Conference on Mobile and Ubiquitous Systems (MobiQuitous'06).* 1–8. DOI:http://dx.doi.org/10.1109/MOBIQW.2006.361781

Sini Ruohomaa, Lea Kutvonen, and Eleni Koutrouli. 2007. Reputation management survey. In *Proceedings of the 2nd International Conference on Availability, Reliability and Security (ARES'07)*. 103–111.

Jordi Sabater. 2004. Toward a test-bed for trust and reputation models. In *Proceedings of the Workshop on Deception, Fraud and Trust in Agent Societies at the 3rd International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'04)*, 101–105.

Jordi Sabater and Carles Sierra. 2001. REGRET: Reputation in gregarious societies. In *Proceedings of the 5th International Conference on Autonomous Agents (AGENTS'01).* ACM, New York, NY, 194–195.

Amirali Salehi-Abari and Tony White. 2012. DART: A distributed analysis of reputation and trust framework. *Computational Intelligence* 28, 4 (November 2012), 642–682. DOI:http://dx.doi.org/10.1111/j.1467-8640.2012.00453.x

Johannes Sänger, Christian Richthammer, and Günther Pernul. 2015. Reusable components for online reputation systems. *Journal of Trust Management* 2, 1 (2015), http://dx.doi.org/10.1186/s40493-015-0015-3

Michael Schillo, Petra Funk, and Michael Rovatsos. 2000. Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence* 14, 8 (2000), 825–848.

Andreas Schlosser, Marco Voss, and Lars Brückner. 2006. On the simulation of global reputation systems. *Journal of Artificial Societies and Social Simulation* 9, 1 (Jan. 2006), page 4.

Emilio Serrano, Michael Rovatsos, and Juan Botia. 2012. A qualitative reputation system for multiagent systems with protocol-based communication. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12),* Vol. 1. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 307–314.

Rob Sherwood, Seungjoon Lee, and Bobby Bhattacharjee. 2006. Cooperative peer groups in NICE. *Computer Networks* 50, 4 (March 2006), 523–544. DOI:http://dx.doi.org/10.1016/j.comnet.2005.07.012

Shanshan Song, Kai Hwang, Runfang Zhou, and Yu-Kwong Kwok. 2005. Trusted P2P transactions with fuzzy reputation aggregation. *IEEE Internet Computing* 9, 6, Special Issue on Security for P2P and Ad Hoc Networks (November 2005), 24–34. DOI:http://dx.doi.org/10.1109/MIC.2005.136

Girish Suryanarayana, Mamadou H. Diallo, Justin R. Erenkrantz, and Richard N. Taylor. 2006. Architectural support for trust models in decentralized applications. In *Proceedings of the 28th International Conference on Software Engineering (ICSE'06)*. ACM, New York, NY, 52–61.

Girish Suryanarayana and Richard N. Taylor. 2006. TREF: A threat-centric comparison framework for decentralized reputation models. University of California, Irvine, CA, ISR Tech. Rep. UCI-ISR-06-2, Jan. 2006.

Girish Suryanarayana and Richard N. Taylor. 2007. SIFT: A simulation framework for analyzing decentralized reputation-based trust models. University of California, Irvine, CA, ISR Tech. Rep. UCI-ISR-07-5, 2007.

Sokratis Vavilis, Milan Petkovic, and Nicola Zannone. 2014. A reference model for reputation systems. *Decision Support Systems* 61, 147–154, 2014.

Kevin Walsh and Emin Gün Sirer. 2005. Fighting peer-to-peer SPAM and decoys with object reputation. In *Proceedings of the 2005 ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems (P2PECON'05)*. ACM, New York, NY, 138–143. DOI: http://doi.acm.org/10.1145/1080192.1080204

Kevin Walsh and Emin Gün Sirer. 2006. Experience with an object reputation system for peer-to-peer filesharing. In *Proceedings of the 3rd Conference on Networked Systems Design & Implementation (NSDI'06)*, Vol. 3. USENIX Association, Berkeley, CA, 1–14.

Wei Wang, Guosun Zeng, and Lulai Yuan. 2006. Ant-based reputation evidence distribution in P2P networks. In *Proceedings of the 5th International Conference on Grid and Cooperative Computing (GCC'06)*. IEEE Computer Society, Washington, DC, 129–132. DOI: http://dx.doi.org/10.1109/GCC.2006.29

Xinlei Wang, Wei Cheng, Prasant Mohapatra, and Tarek F. Abdelzaher. 2013. ARTSense: Anonymous reputation and trust in participatory sensing. In *Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM'13)*. IEEE Computer Society, 2517–2525. DOI: http://dx.doi.org/10.1109/INFCOM.2013.6567058

Andrew G. West, Sampath Kannan, Insup Lee, and Oleg Sokolsky. 2010. An evaluation framework for reputation management systems. In *Trust Modeling and Management in Digital Environments: From Social Concept to System Development* (Zheng Yan, ed.), 282–308. Information Science Reference, Hershey, PA.

Li Xiong and Ling Liu. 2004. PeerTrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Transactions on Knowledge and Data Engineering* 16, 7 (July 2004), 843–857. DOI: http://dx.doi.org/10.1109/TKDE.2004.1318566

Zheng Yan, Yu Chen, and Yue Shen. 2013. A practical reputation system for pervasive social chatting. *Journal of Computer Systems Science* 79, 5 (August 2013), 556–572. DOI: http://dx.doi.org/10.1016/j.jcss.2012.11.003

Bin Yu and Munindar P. Singh. 2002. An evidential model of distributed reputation management. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1 (AAMAS'02)*. ACM, New York, NY, 294–301. DOI: http://doi.acm.org/10.1145/544741.544809.

Bin Yu and Munindar P. Singh. 2003. Detecting deception in reputation management. In *Proceedings of the 2nd International Joint Conference on Autonomous agents and Multiagent Systems (AAMAS'03)*, Melbourne, Australia, July 2003. ACM, New York, NY, 73–80. DOI: http://doi.acm.org/10.1145/860575.860588

Giorgos Zacharia and Pattie Maes. 2000. Trust management through reputation mechanisms. *Applied Artificial Intelligence* 14, 9 (2000), 881–907.

Yulian Zhang and Lihua Wang. 2012. A new reputation mechanism based on referral's credibility for p2p networks. In *Proceedings of the 11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science (DCABES'12)*. IEEE Computer Society, Washington, DC, 153–156. DOI: http://dx.doi.org/10.1109/DCABES.2012.60

Yan Zhang, Wei Wang, and Shunying Lü. 2007. Simulating trust overlay in P2P networks. In *Proceedings of the 7th International Conference on Computational Science, Part I*: (ICCS'07), Yong Shi, Geert Dick Albada, Jack Dongarra, and Peter M. Sloot (Eds.). Springer-Verlag, Berlin, 632–639. DOI: http://dx.doi.org/10.1007/978-3-540-72584-8_84

Runfang Zhou and Kai Hwang. 2007. PowerTrust: A robust and scalable reputation system for trusted peer-to-peer computing. *IEEE Transactions on Parallel Distributed Systems* 18, 4 (April 2007), 460–473.

Received April 2015; revised July 2015; accepted September 2015