Journal of Information Systems Engineering and Management
2023, 8(1), 19218
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Literature Review

Reputation Systems: A framework for attacks and frauds classification

Rui Humberto Pereira[1]*, Maria José Angélico Gonçalves[2], Marta Alexandra Guerra Magalhães Coelho[3]

[1] CEOS.PP/ University of Maia, 4475-690 Maia, Porto, Portugal
[2] CEOS.PP, ISCAP, Polytechnic of Porto, 4465-004 Mamede Infesta, Porto, Portugal
[3] ISCAP, Polytechnic of Porto, 4465-004 Mamede Infesta, Porto, Portugal
***Corresponding Author:** rhpereira@umaia.pt

Citation: Pereira, R. H., Gonçalves, M. J., & Magalhães, M. A. G. (2023). Reputation Systems: A framework for attacks and frauds classification. *Journal of Information Systems Engineering and Management, 8*(1), 19218. https://doi.org/10.55267/iadt.07.12830

ARTICLE INFO

Received: 05 Dec 2022
Accepted: 13 Jan 2023

ABSTRACT

Reputation and recommending systems have been widely used in e-commerce, as well as online collaborative networks, P2P networks and many other contexts, in order to provide trust to the participants involved in the online interaction. Based on a reputation score, the e-commerce user feels a sense of security, leading the person to trust or not when buying or selling. However, these systems may give the user a false sense of security due to their gaps. This article discusses the limitations of the current reputation systems in terms of models to determine the reputation score of the users. We intend to contribute to the knowledge in this field by providing a systematic overview of the main types of attack and fraud found in those systems, proposing a novel framework of classification based on a matrix of attributes. We believe such a framework could help analyse new types of attacks and fraud. Our work was based on a systematic literature review methodology.

Keywords: e-commerce, trust, reputation systems.

INTRODUCTION

Commercial transactions require that the participants trust each other. This trust gives participants a notion about the risk, thus, leading the person to conclude, or not, the transaction. In e-commerce, this sense of risk/security based on trust is much more critical, particularly when there is no prior knowledge of the person on the other side. In e-commerce, the users must be aware of some interrelated aspects regarding the other participant in the transaction: (1) the real identity; (2) honesty; and (3) the quality of the product/service. Regarding the real identity of the person, this can be a complex problem in on-online marketplaces, due to the incorrect, or absence, of an effective identity validation of the person associated with the user (profiles of buyer, seller, or both) in that e-commerce platform. On the other hand, regarding the honesty of the user and the quality of the product/service, these issues have been addressed by means of reputation and recommendation systems. In the case of B2C, when there is a company or brand associated with the platform, the buyers' trust is mainly based on their prior knowledge about the credibility of that company, brand, or product/service quality, which may be additionally complemented by a reputation and recommendation system.

Reputation and trust are distinct and interrelated concepts. Jøsang *et al.* (2007) distinguish "Reliability trust", and "Decision trust". In the former concept, the author uses the definition proposed in (Gambetta, 1988). However, the authors consider the concept of trust to be more complex, referring to this as such: *Decision trust* as: "*Trust is the extent to which one party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible*" (Jøsang et al., 2007, p. 620). Regarding to the concept of reputation, the same authors define reputation according to the Concise Oxford Dictionary, as: "*Reputation is what is generally said or believed about a person's or thing's character or standing*" (Jøsang et al., 2007, p. 620). Thus, during an e-commerce transaction, trust and reputation are two subjective concepts on which the decision to conclude is based, accepting a certain level of risk. This observation lets us identify the first limitation of reputation systems.

As previously mentioned, the reputation of a product or user can be determined by means of reputation systems. In the present work, we focus on the reputation of the users, since it is a distinct problem of product reputation/recommendation, despite the fact that they share common principles. We should notice that there are other types of threats, such as the ones in the field of cybersecurity, at infrastructure and network levels, that we do not consider in this work, because they are distinct from the reputation systems as a mean to provide trust to the users.

Reputation systems have been widely used in peer-to-peer (P2P) networks for establishing a user's reputation score based on what he gives to the network and gets, in terms of the criteria for choosing files (Damiani et al., 2002). Crowdsourcing platforms also could apply similar principles, but in this case, it is to establish a rating score of the user's reputation in terms of the value of tasks (Gong et al., 2021). In theory, all collaborative network environments could benefit from reputation systems as a mean to provide trust. In the context of our work, e-commerce marketplaces, such as Amazon and eBay, also apply reputation systems in order to enable their users to rate other users. This is to say, typically a buyer rates a seller after a finished transaction. Our focus will be on this latter case, in which the trust of an e-commerce user, at the moment of deciding about a transaction, is based on the reputation of the other participant.

We expect to contribute to the increase of knowledge in the field, by answering the following research question:

What has been researched about the most common attacks and frauds on e-commerce platforms that may affect the user's trust based on reputation systems?

In order to answer this question, we choose a systematic literature review as methodology using three citations databases: Web of Science, Scopus, and Google Scholar. Next, we proposed a new framework for attacks and fraud classification. We believe that such a framework can be a useful tool for analysing new types of attacks and fraud, thus, contributing to the knowledge in the field.

In the next section, we present the concepts related to reputation systems focusing our discussion on e-commerce communities. In the section Methodology, we present the methodology used, based on a literature review. The discussion, about the types of attacks, and frauds and how these security issues are classified in the literature, is presented in the section Discussion. After, we discuss our proposal for a new framework of classification and perform its validation based on the scenarios found in the literature. We finish the paper presenting our final remarks and the focus of our future work.

BACKGROUND ON REPUTATION SYSTEMS

In order to contextualize the reader as to the weaknesses of reputation systems, we consider it important to clarify some concepts, principles and strategies that are usually adopted in these systems.

The scope of the present paper is reputation systems for e-commerce users, however, the discussion of the following aspects and principles is generalizable to other types of collaborative networks, as well as to product recommendation systems in e-commerce. Jøsang et al. (2007) distinguish reputation systems from recommendation systems referring to them as collaborative sanctioning and collaborative filtering. In reputation systems (i.e., collaborative sanctioning), the user is judged after a transaction. This contrasts with the recommendation systems (i.e., collaborative filtering), which are based on different tastes and the subjective opinions of the users.

Hendrikx et al. (2015) propose a taxonomy for reputation systems, which, at the first, classifies the reputation systems as implicit and explicit. According to the authors, implicit reputation systems are systems that do not have a defined reputation system, although reputation information is used by its members to assist in decision making. Examples of such reputation approaches exist in social networks (e.g., Facebook and LinkedIn), in which we can extract some degree of trust from the information gathered through friends of friends. Another example is Google's search engine, in which the order of the search results represents a ranking of pages, based on the reputation of each page. The reputation is determined by the number of links that point to the page, and where the links originate (Hendrikx et al., 2015). On the other hand, explicit reputation systems have implemented a model that enables the estimate of a reputation using a score. The latter are the focus of the present paper.

The reputation estimation model encompasses the three dimensions (1) sources and types, of data, (2) the algorithm based on mathematical calculations and (3) the type of output for the reputation score and how it is disseminated. In (Hoffman et al., 2009), the authors refer to these three dimensions as Formulation, Calculation and Dissemination. The accuracy of the reputation score depends on the effectiveness of the model, as well as the types of threats that it is immune to. Another characteristic of the model is its architecture, which is a central or distributed system. On the second level of the taxonomy proposed by Hendrikx et al. (2015) the identified aspects are systematized, as well as discussed in other works (Hoffman et al., 2009). In the following section, we will briefly discuss these aspects.

Sources and type of data

The sources of information that support the formulation of reputation provide the raw data that feeds the algorithm implemented at the computational level. This data is diverse and complementary to each other. We can group it into two main sources: Manual and automatic. Hoffman et al. (2009) suggest the following classification of sources of information:

- Manual sources are obtained from human feedback, usually in the form of user ratings of other identities based on the experience of a single transaction such as the feedback in a marketplace, a specific time period or arbitrary feedback;
- Automatic sources are obtained automatically either via direct or indirect observation.
    - Direct observations provide data regarding directly observed events such as the success or failure of

interaction, the direct observations of cheating, or in the case of the P2P network, the measurement of resource utilization by neighbours;

- Indirect observations are obtained second-hand or are inferred from first-hand information.

These sources of information could influence the reputation positively, negatively, or neutrally, according to its level of relevance, calculated by the algorithm implemented in the system. Regarding to the taxonomy of the datatypes, this data can be binary, discrete or continuous. Other types are possible but could restrict the type and accuracy of the results of the computational algorithm. In order to achieve a quantitative metric on user reputation, the qualitative-input datatypes could require the conversion to a quantitative value. For example, free text reviews can complement a numerical score but require some manual analyses, which are not viable, or processed by means of artificial intelligent mechanisms in order to convert to a quantitative variable.

The computational approach applied to calculate the reputation value of the target

The result of the algorithm that computes the data, obtained from manual or automatic data sources, consists of a metric regarding the reputation of a user, which in general is a quantitative one. Several algorithm-based approaches for reputation models can be found in the literature (C. Dellarocas, 2000; Hendrikx et al., 2015; Panagopoulos et al., 2017). The main challenge placed on the designer of such algorithms is choosing which are the input variables and their respective weights for the output metric.

The temporal variable is another factor that some models consider in their mathematical analyses, in which the impact (or relevance) of the feedback, or direct/indirect observation, decreases with time, which is data ageing (Hendrikx et al., 2015).

Regarding new users, who do not have historical records of transactions, and for which an initial reputation score can be estimated, default neutral value for reputation is typically set forth. Panagopoulos et al. (2017) discuss some approaches for dealing with newcomers, as well as other economic and social issues such as inducing user participation, using incentives, and dealing with reciprocity and retaliation.

Other approaches such as the ones based on machine learning (Wang et al., 2020), for automatic detection of false or unfair ratings, or others blockchain-based (Zulfiqar et al., 2021) approaches, in which the financial model is not viable for a dishonest user. Furthermore, in distributed architectures reputation data is shared among several e-commerce platforms. Below, in the subsection 0, we detail this discussion.

The output reputation score can be classified as either binary, discrete, or continuous. A binary one could represent if the user is reputable or not. The discrete outputs, for example, one to five stars, define the level of reputation, as well as the continuous scores, but in this case, give a much fine-grained classification.

Accuracy and immunity of the reputation model

The accuracy of the model depends on the quality of the input data and the robustness of the algorithm and mathematical approach. Additionally, the accuracy can be subject to fraud and manipulation. These threats to the reputation systems have two possible purposes: to increase or decrease the reputation of a user, based on a malicious strategy (Koutrouli & Tsalgatidou, 2012). In the Discussion Section, we will examine these threats in detail, which is the focus of the present paper.

The incentives for participation in rating the transaction are one approach to increase the volume of input data, which is important to get reliable outputs. However, Panagopoulos et al. (2017) claim that, although user participation is necessary for successful feedback-based reputation systems, most e-commerce communities do not provide any kind of incentives to encourage it. That is due to the fact that e-commerce platforms usually achieve good enough participation through the mutual exchange of ratings between the members involved in the transaction, which takes place right after its completion, by courtesy. The aforementioned approaches, based on machine learning and public blockchain networks, could help to mitigate these problems.

Centralized vs distributed architecture

In reputation systems based on a centralized architecture, the data is managed only by one entity. If a user has two accounts, each in a distinct e-commerce platform based on centralized reputation systems, then he has two profiles, each one with its own reputation score, perhaps two incoherent values of reputation. Several proposals for decentralization can be found in the literature, but other similar problems emerge (Panagopoulos et al., 2017). In recent literature (Ahn et al., 2018, 2019; Dennis & Owen, 2015; Dhakal et al., 2019; Karode et al., 2020; Moher et al., 2009; Schaub et al., 2016; Zeynalvand et al., 2021; Zulfiqar et al., 2021), blockchain-based approaches are proposed to enable a distributed architecture in reputation systems in terms of sharing data. According to the authors, these approaches ensure transparency and could help mitigate some types of known fraud.

Zulfiqar et al. (2021) state that the central authorities can potentially filter, tamper, add, or reject product reviews based on their preference. Schaub et al. (2016) state that, potentially, a centralized system can be abused by the central authority.

The management of users' reputation, based on payment systems, are also prone to manipulation by malicious entities, which include the advertisers or owners themselves, who may give extremely high or low ratings on purpose (Ahn et al., 2019; Dennis & Owen, 2015).

Dhakal and Cui (2019) present the same arguments, stating that the current centralized systems are silos and not transparent in the review process. Besides the lack of transparency, these isolated centralized systems do not benefit from the reputation data of each other. Zeynalvand et al. (2021) state that it is hard to derive trust models that are robust to attacks such as whitewashing and Sybil attacks, if users do not share information.

Karode et al. (2020), in the context of travel review systems, state that blockchain-based reputation systems enable consumers to be confident that the review score is not affected by the platform providers. Besides, the businesses can maintain the same rating score regardless of the platform they take part in. Low-quality review handling is a challenge for the global-scale review system; however, this problem can be addressed with automatic filtration.

METHODOLOGY

Our review can be categorized as a systematic review of the scientific literature on security problems in user reputation systems.

Systematic reviews are a form of meta-analysis designed to collect, investigate, and summarise what is known and what is not known about a "specific practice-related question" (Briner et al., 2009). Systematic reviews are used across a broad range of disciplines. Qualitative studies have established a place for themselves within the methodologies, as evidenced by initiatives such as the Cochrane qualitative methods group (Dixon-Woods & Fitzpatrick, 2001) and textbooks such as Systematic Reviews in the Social Sciences (Petticrew & Roberts, 2005) and An Introduction to Systematic Reviews (Thomas et al., 2017).

In this study, besides conducting the literature review following its primary objectives according to Moher (2009) we also substantiate the results obtained with a literature review, presenting theoretical perspectives and innovations from leading authors in the field. According to the authors, the systematic literature review is carried out in 3 steps (Moher et al., 2009). First, the research question is defined; this is followed by a research protocol for evaluating the selected scientific articles. The last step involves answering the research questions (in the first step), based on the scientific articles identified as relevant (in the second step). **Figure 1** summarizes the steps followed by the adopted methodology.

The first step of the adopted methodology is related to the definition of the research question of this study. The main research question intends to identify the state of the art concerning our study characteristics. Therefore, our research question can be formulated as follows, as above mentioned in the introduction section: *What has been researched about the most common attacks and frauds on e-commerce platforms that may affect the user's trust based on reputation systems*?

After the definition of the research question, the second step was related to the selection of the empirical data to be analysed. Data collection took place in October 2022. We did not apply any chronological filter. In the first phase, we tried a separate search for each keyword. In Web of Science Core Collection (WOS) we applied the following strategy: Search: (TITLE-ABS-KEY("reputation system" and taxonomy and attack) OR TITLE-ABS-KEY("reputation system" and classification and attack) OR TITLE-ABS-KEY("reputation system" AND type attacks)). In SCOPUS and Google Scholar (GS) we followed the same criteria. This search resulted in 38 articles selected from WoS, 75 articles selected from Scopus and 29 selected from GS. The lists were exported to excel for further analysis, and the following fields were chosen: authors, title, year, link, abstract, and keywords.

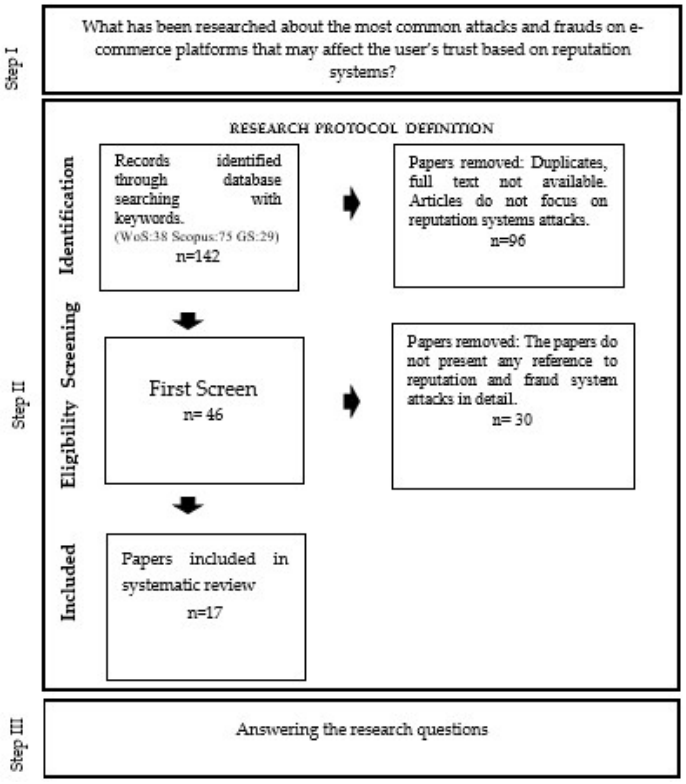

Figure 1. Systematic review structure

Then, we evaluated the articles based on the criteria for inclusion to determine their relevance to the study. An article had to include the search terms as the core technology under analysis. This was typically demonstrated by its title, abstract, and keywords. We only selected academic peer-reviewed journal articles and conference proceedings and excluded others, namely: (a) articles no fully available, (b) articles not available in English, (c) duplicate articles, and (d) articles that did not discuss security issues in user reputation systems. Our initial search was carried out in October 2022 and yielded 142 articles. Once we eliminated duplicates, we were left with a population of 96 articles. After, our research team, including a professor and a master's student, reviewed this collection of articles for relevancy. In the first round, we assessed the articles for relevance based on title, abstract, and keywords. This process led to the selection of 46 articles. Any articles we did not agree upon were also excluded. In the second round of revisions, we assessed the articles based on the full paper. We eliminated 29 articles that did not present detailed reference to attacks and frauds on the reputation system. Thus, we identified 17 relevant articles for analysis, which are listed in **Table 1**.

Table 1. Selected Articles

Reference	Article Title
(C. Dellarocas, 2000)	Immunizing online reputation reporting systems against unfair ratings and discriminatory behaviour
(C. Dellarocas., 2000)	Mechanisms for coping with unfair ratings and discrimination behaviour in online reputation reporting Systems
(Douceur, 2002)	The Sybil Attack. In: Druschel, P., Kaashoek, F., Rowstron, A. (eds) Peer-to-Peer Systems.
(C. N. Dellarocas, 2003)	The Digitization of Word-of-Mouth: Promise and Challenges of Online Feedback Mechanisms
(Jøsang et al., 2007)	A survey of trust and reputation systems for online service provision
(Hoffman et al., 2009)	A survey of attack and defense techniques for reputation systems
(Swamynathan et al., 2010)	The design of a reliable reputation system
(Fraga et al., 2012)	A Taxonomy of Trust and Reputation System Attacks
(Koutrouli & Tsalgatidou, 2012)	Taxonomy of attacks and defense mechanisms in P2P reputation systems— Lessons for reputation system designers
(Feng et al., 2012)	Vulnerabilities and countermeasures in context-aware social rating services
(Yao et al., 2012)	Addressing Common Vulnerabilities of Reputation Systems for Electronic Commerce
(Sänger et al., 2015)	Reusable components for online reputation systems
(Koutrouli & Tsalgatidou, 2016)	Reputation Systems Evaluation Survey
(Panagopoulos et al., 2017)	Modeling and Evaluating a Robust Feedback-Based Reputation System for E-Commerce Platforms
(Camilo et al., 2020)	A Secure Personal-Data Trading System Based on Blockchain, Trust, and Reputation
(Zulfiqar et al., 2021)	EthReview: An Ethereum-based Product Review System for Mitigating Rating Frauds
(Zeynalvand et al., 2021)	A Blockchain-Enabled Quantitative Approach to Trust and Reputation Management with Sparse Evidence

DISCUSSION

Types of vulnerabilities and attacks

Dellarocas (2000, 2000, 2003) focused his work on the fraudulent behaviour of users when rating others, in online trading communities. The author has identified two scenarios: (1) unfair buyer ratings and (2) discriminatory seller behaviour. In the case of unfair buyer ratings, there are two scenarios: (1.a) Unfairly high ratings ("ballot stuffing") and (1.b) Unfairly low ratings ("bad-mouthing"). In both cases, a seller colludes with a group of buyers, but in the first case, "ballot stuffing" to increase his own reputation and, for "bad-mouthing" to damage the reputation of other sellers, his competitors. This can, potentially, increase his orders and decrease the orders of his competitors. The second group of scenarios relates to discriminatory seller behaviour: (2. a) harmful discrimination - a seller provides good service to the majority of buyers, except a few specific ones that they "don't like". This kind of action does not have a great impact on the seller's reputation if the number of "victims" is small; (2.b) positive discrimination – In this strategic action, the seller can potentially increase his own reputation by providing an exceptional quality service to a few buyers and average quality service to the rest of the buyers. If the number of privileged customers is sufficiently large, this action is equivalent to Ballot stuffing, but in an imperceptible way, without having to conspire with third parties.

Douceur (2002), in the context of P2P networks, following the suggestion of Brian Zill, inspired by a book[1] dated 1973 with the same name, coined the term Sybil attack. In this type of attack on network services, an entity forges multiple identities in the system in order to increase his influence. In the context of e-commerce, the typical scenario is an entity that can influence the reputation of a user, or product by using multiple identities to insert unfair ratings into the system. A similar type of attack, also based on a lack of effective identity management, is Whitewashing. In this type of attack, a dishonest buyer is able to whitewash its low trustworthiness by starting a new account with the initial trustworthiness value or using some vulnerability in the system. In (Fraga et al., 2012), the authors distinguish between Re-entry and Whitewashing. According to the authors, in the former case, if attackers can create new "identities" freely, this presents the opportunity to remove a bad reputation by creating a new identity. In Whitewashing, the attackers can repair their reputation completely by using some system vulnerability. In our opinion, Re-entry is a specific case of a Whitewashing attack, in which the malicious user uses vulnerabilities in identity management in order to get a new identity with a clean reputation. Despite their distinct purposes, Sybil, Whitewashing or Re-entry attacks are based on limitations in the identity management mechanism mainly due to low effort, or low cost, to get a new identity. The centralized nature of the reputation systems, used in e-commerce platforms that don't share the user identity data, inflates the problem (Camilo et al., 2020; Zulfiqar et al., 2021)

In Swamynathan et al. (2010), the authors define the Churn attack in the context of P2P networks as high rates of peer turnover. In such a scenario, a significant number of peers will have relatively short-term accumulated reputation scores as a result of a small number of past interactions. In e-commerce platforms, this undesirable scenario can be enabled by the low effort/cost to get several identities.

The Sybil attacks are frequently coupled with Collusion attacks, in which the different identities of a single entity conduct coordinated actions in order to influence the reputation of a user or product. Koutrouli and Tsalgatidou (2012) identify three variants of Collusion: (1) collusive badmouthing and (2) collusive reducing recommendation reputation and (3) Collusive deceit.

As we can observe, all these types of attacks on reputation systems result from vulnerabilities in one component of the underlying reputation model or architecture of the system. The aforementioned problems of lack of transparency in the management of the reputation data, discussed in the section

[1] https://en.wikipedia.org/wiki/Sybil_(Schreiber_book)

Background on reputation systems, subsection 0, the distinct models and criteria to calculate the users' reputation score, as well as the centralized nature of the current reputation systems make the mitigation of those attacks a very difficult task.

In the literature, one can find several taxonomies and categorizations of these types of attacks. In the next section, we discuss our findings in terms of taxonomies and classifications of attacks that derived from our systematic literature review.

Taxonomies of attacks

Feng et al. (2012) discussed three types of attacks: direct, disguise, and misguidance attacks. A direct attack is common in most basic rating systems. The attackers provide dishonest ratings only on the items within the attack target set. Then all user ratings for an item are computed to obtain its aggregate recommendation score. Thus, the social recommendation score will reflect users' mainstream opinions of. On the other hand, disguise attacks and misguidance attacks are representative of trust-enhanced social rating systems. According to the authors' definition, in these reputation systems, the user trust can be defined as a credibility measurement of the rating. Thus, in these two types of attacks sophisticated strategies of non-direct rating are applied in order to reduce the credibility of honest users while increasing the credibility of the malicious ones. Self-promoting is an example of a disguise attack. In the case of a misguidance attack, the strategy aims at strategically making the system misjudge the honest rating behaviour to be dishonest and the dishonest rating behaviour to be honest. We classified the authors' proposal as simplistic, because each category of attack is very comprehensive, in which many different types of attack can be placed.

Fraga et al. (2012) proposed an attack taxonomy based on the reputation system architectural model and a set of well-known security topics. In one dimension, the authors propose three basic processes: trust and reputation (T&R) information gathering, T&R calculation, and T&R dissemination. In the second dimension, the authors propose Primary topics: Authentication, Authorization, Availability and Utility/process; and Derived topics: Identification, Non-repudiation, Confidentiality, Integrity and Time Integrity. For instance, bad-mouthing and ballot-stuffing regards T&R information gathering and Utility/process primary topic. According to the authors, Utility means usefulness. It means that the action an agent wants to perform over a resource could be carried out correctly. In this example of fraud, the resource is the users' ratings. The authors aim to propose a holistic taxonomy that may include all known attacks and provides structured tools to identify new attacks. In our opinion, the proposed framework is very complex, and not compliant with the comprehensibility requirement of a good taxonomy, as stated by the authors.

Yao et al. (2012) distinguish the vulnerabilities of reputation systems into two categories: (1) system-based vulnerabilities – that relate to the foundation and environment of reputation systems, and (2) metric-based vulnerabilities that tie to the selected reputation metric and its updates. These two categories encompass six security requirements for the reputation system, at a lower level, and at a higher level of computation of reputation scores and decision-making, i.e., the reputation metric. In the first category, the authors include vulnerabilities in the message exchange between system nodes, fragilities in the system nodes in terms of data integrity, when storing and processing it, and the lack of effective identity management that avoids malicious actions triggered by multiple identities of the same user. The second category relates to the fragilities in the model in which the reputation scores are based on decision-making.

Koutrouli and Tsalgatidou (2012, 2016), in the context of P2P networks, and Panagopoulos et al. (2017), for e-commerce, propose the following taxonomy of types of attack: (1) Strategy-Based Attacks, (2) Identity-Based Attacks, and (3) Unfair Ratings. This proposal strictly focuses on the reputation model on which the metrics are based. This taxonomy proposal includes the types of attacks and their variants known in the literature. However, despite its coverage, we observe that other possible combinations of attack could potentially exist but are not included in this taxonomy, such as reciprocity and retaliation, as discussed by the same authors in another paper (Panagopoulos et al., 2017).

Panagopoulos et al. (2017) also refer to other issues from the economic and/or social perspectives, which are not attacks but may affect the effectiveness of the reputation systems as a means to provide trust in the e-commerce community. The authors refer to (1) the importance of inducing user participation using incentives, as the average quantity of feed-backs is insufficient to get an accurate score of the reputation of a person, (2) dealing with reciprocity (the authors mention studies where a strong correlation between buyer and seller ratings are identified), and retaliation, and (3) how to deal with the reputation of newcomers, in terms initial reputation for these users.

Additionally, Panagopoulos et al. (2017) still refer to a number of issues found in decentralized reputation systems, such as trust propagation (i.e., how to effectively communicate trust information in large-scale networks of loosely connected entities (Zeynalvand et al., 2021), and storage of local and global reputation information.

Sänger et al. (2015) propose a different approach of taxonomy of attacks. The authors propose, at the highest level, to distinguish between seller attacks and advisor attacks. In these major classes, the authors classify every type of attack into two dimensions: attackers and behaviour. At a lower level, the attacker's dimension refers to the number and characteristics of the digital identities participating in an attack (one identity, multiple identities or multiple entities), which distinguishes Sybil and Collusion, and behaviour dimension (consistent or inconsistent). The authors claim that they are focused on the general characteristics and symptomatology of attacks, such as the continuity and the number of attackers. We consider this approach for a taxonomy very compact and at the same time comprehensive, which makes it very interesting. However, approach can be limiting in terms of effective identification of the type of attack. For instance, we do not see how it is possible to classify the whitewashing attack, and the authors, do not provide examples or additional information. Another example is the scenario of a distributed reputation system in which the reputation data is manipulated.

Hoffman et al (2009) classify attacks against reputation systems based on the goals of the reputation systems targeted by these attacks. The authors propose five categories of attack: (1) Self-Promoting, (2) Whitewashing, (3) Slandering, (4) Orchestrated and (5) Denial of Service (DoS). We consider that this classification includes vulnerabilities at two levels: at a higher level, in the reputation model, which includes the first four categories, and at a lower level, the fifth category that relates to the DoS attacks. We consider this classification incomplete, namely at the lower level, because DoS attacks are just one of the several possible types at this level, in this case, at the network level. We note that there are many other types of vulnerabilities at a low level. Below, in the present section, we will discuss our proposal for a classification in which we refer to those other possible low-level attacks. Another aspect of this proposal is its focus on the attack's goal. For example, regarding the Whitewashing attack, we ask if the final goal is to escape from the consequences of a low reputation or to manipulate someone's reputation score.

In **Table 2** we summarize our findings in terms of classifications, presenting a brief description of the structure and its focus.

Table 2. Summary of proposed classifications

Reference	Structure	Focus
(Feng et al., 2012)	Three types of attack	e-commerce and collaborative networks
(Fraga et al., 2012)	Bi-dimensional framework: basic processes and topics	Reputation systems in general
(Koutrouli & Tsalgatidou, 2012, 2016; Panagopoulos et al., 2017)	Hierarchical classification of type and variants	P2P and e-commerce
(Sänger et al., 2015)	At highest level: seller attacks and advisor attacks. In these major classes: two dimensions: attackers and behaviour.	Electronic marketplaces
(Hoffman et al., 2009)	Five types of attack	Reputation systems in general

Limitations of the current approaches of categorization

In the present work, an attack is an intentional action with fraud as its objective, which affect the reputation systems. Additionally, other issues can compromise the accuracy of reputation scores; however, they are not attacks. In (Jøsang et al., 2007) the "Bias towards positive rating" is explained as positive ratings simply representing an exchange of courtesies; either the positive rating is given in the hope of getting a positive rating in return, or the negative rating is avoided due to fear of retaliation from the other party.

We consider that there is a gap of extensibility in the proposed classifications found in the literature, as it is difficult to classify all variants of a type of attack because a slight variation can change the given classification. In the next section, we will discuss these limitations compared to our proposal.

PROPOSAL FOR A NEW FRAMEWORK OF CLASSIFICATION

Types of fraud and attacks in reputation models

Analysing the types of fraud and attacks to the models used in reputation systems, founded on the literature, we observed that there are two fundamental levels of attack, of distinct nature (origin and technique), as well as the field of research. The former category regards all vulnerabilities at the network, infrastructure and application levels. These may result from bad definitions in the systems, security breaches in the software, wrong choices in terms of network architecture, missing defence tools, such as firewalls, WAFs, IDS, lack of cryptography in the stored data or messaging exchanging, among others. On the other hand, the second category encompasses the fragilities of the model that defines the algorithm, which gathers and calculates all metrics in order to establish a reputation score, as well as architectural issues, such as centralized vs distributed or identity management. Moreover, in general, these attacks at the model level are accomplished by members of the e-commerce community, e.g., a user which has been registered for a long time in the Amazon or eBay marketplace, contrasting with the attacks at the levels of the network, infrastructure or application, which, in general, are perpetrated by outsiders.

Due to the nature of the attacks, its first category is out of the scope of our work. In the present work, we focused on the vulnerabilities of the model and the architectural issues on which the reputation formulation is based. Thus, we will continue our discussion focused on the vulnerabilities of the reputation model.

We start our discussion by observing that any malicious action falls into one of two fraud cases: to increase or decrease the reputation of an entity or product/service. For instance, ballot stuffing and bad-mouthing are the same vulnerability, but with a different type of fraud as a goal.

In another observation, we notice that many attacks are combinations of primitive types of attack leading to several variations of the same type of attack. For instance, a group of entities can collude by means of unfair ratings (collusion + unfair ratings), or the collusion of several identities, of the same entity, combined in order to give unfair rations, i.e., a Sybil attack.

These observations led us to propose a novel approach based on a classification of multidimensional attributes.

Proposal for a matrix of attributes

Our proposal is based on a classification of multidimensional attributes. Each type of attack has the following five attributes, each with several possible values:

1) Type of fraud: (a) Increase its own reputation, (b) decrease the reputation of others, (c) increase the reputation of others or (d) decrease recommendations reputation of a user.

2) Level: (a) Identity management, (b) model formulation/calculation (Hoffman et al., 2009), (c) formulation/data or (d) architectural – This attribute identifies an element in the model where the vulnerability is.
   - Sybil and whitewashing are examples of an identity management-level attack.
   - The lack of an ageing mechanism, in the model formulation, can be exploited. At the model formulation level, we also consider the lack of validation (policing) of the ratings by means of a manual (endorsers) or automatic mechanisms (e.g. machine learning);
   - In terms of scenarios of architectural issues, in a centralized system, the entity that manages the platform can manipulate the reputation data. In the case of decentralized systems, the nodes can potentially manipulate the data shared in the network, even if the data is encrypted or signed.
3) Cardinality – (a) One entity, (b) multiple entities; (c) multiple identities or (d) Many entities to many entities.
   - In a Sybil attack an entity with multiple identities participates;
   - In the case of Collusion, the attack is performed by multiple entities.
4) Behaviour – (a) One time, (b) constant or (c) variable
   - The scenario of unfair ratings given by one, or more buyers to sellers in a constant or variable behaviour;
   - Following a variable pattern, e.g. Oscillatory (Panagopoulos et al., 2017) or Traitors (Panagopoulos et al., 2017) attacks.
5) Action – (a) Unfair rating, (b) discriminatory rating, (c) creating a new identity or (d) data manipulation

These five dimensions proposed to classify the possible attacks on the reputation systems are focused on e-commerce communities. However, this matrix could potentially be applied to other network services, such as crowdsourcing or P2P.

The proposed approach also has the advantage of extensibility. That is to say, new attributes can be added to the framework, as well as new values for these attributes. Additionally, our approach can handle multiple variants of the same type of attack, avoiding long and complex hierarchical taxonomies. This is a substantial advantage regarding the taxonomies found in the literature. For instance, one can perform bad-mouthing fraud applying collusion by means of Sybil, or not. Thus, we consider bad mounting as a type of fraud of the attack, lack of effective identity management at the level in the case of a Sybil attack and cardinality as multiple identities, in the case of a collusion-based attack.

Evaluation of the proposed matrix

In this section, we will test our proposal on the types of attacks on reputation systems found in the literature. In **Table 3**, for each attack (first column) we present all five attributes. When all attributes are possible, we use "any". In such cases, it means that the same attack (or variant) has variants, as many as the number of possible combinations of attributes.

Table 3. Attacks multidimensional analyse

Attack	Description	Fraud	Level	Cardinality	Behaviour	Action
Ballot stuffing	(C. Dellarocas., 2000; C. Dellarocas, 2000; C. N. Dellarocas, 2003)	(a)	(b)	Any	Any	(a), (b) or (d)
Bad-mouthing		(b)	(b)	Any	Any	(a), (b) or (d)
On-off	(Alshammari et al., 2021)	Any	(b)	Any	(c)	(a) or (b)
Oscillatory behaviour	(Panagopoulos et al., 2017)	Any	(b)	Any	(c)	(a) or (b)
Quality variations over time	(Jøsang et al., 2007)	Any	(b)	Any	(c)	(a) or (b)
Sybil	(Douceur, 2002)	Any	(a)	(c)	Any	(c) and ((a) or (b))
Whitewashing	(Fraga et al., 2012)	(a)	(a), (b) or (c)	(a)	(a)	Any
Re-entry	(Fraga et al., 2012)	(a)	(a)	(a)	(a)	(c)
Churn	(Swamynathan et al., 2010)	(a)	(a)	(a)	(a)	(c)
Collusion	(Koutrouli & Tsalgatidou, 2012, 2016)	Any	(b)	(b) or (c)	(b) or (c)	(a) and/or (b)
Collusive deceit		(b) and (c)	(b)	(d)	(b) or (c)	(a)
Collusive badmouthing		(b)	(b)	(b) or (c)	(b) or (c)	(a)
Collusive reducing recommendation reputation		(d)	(b)	(b) or (c)	(b) or (c)	(a)
Data manipulation by a central authority	(Schaub et al., 2016; Zulfiqar et al., 2021)	Any	(c)	(a)	Any	(d)
Reputation Trap	(Feng et al., 2012)	(b) and (c)	(b)	(d)	(b) or (c)	(a) and (b)
Bias toward positive rating	(Jøsang et al., 2007)	(c)	(b)	(a)	(a)	(a)

In **Table 3**, we can observe that, only by the attack name, one cannot know all the details of the malicious action. For instance, a malicious user may apply a Sybil attack to increase his own reputation or damage someone's reputation. Thus, our proposal enables security analysts to classify new types of attack, as well as identify two types of attack as effectively being the same, for instance: On-off and conflicting-behaviour attacks. Even if two types of attack have distinct proposals, if their attributes are the same, then, potentially, the same approach for dealing with them could be applied in both cases.

CONCLUSION

In the present work, we conducted a systematic literature review in order to systematize the several types of attacks and fraud to reputation systems in the context of user reputation in e-commerce. In our discussion, we present some observations that lead us to conclude that the type of attack/fraud does not inform us about all the necessary details for understanding the malicious action. In fact, each vulnerability may be combined with others. Thus, the same type of attack could have distinct names, or many variants, making it very difficult to inbox it in a hierarchical or group-based classification, as the ones found in the literature.

In order to overcome this gap, we propose a novel framework of classification. We are convinced that our approach has advantages over other proposals based on taxonomies, categories or hierarchical classifications, which are complex and redundant when trying to cover all types/variants. We expect to contribute to the knowledge in this research field by means of our proposal of an innovative framework. We believe that our framework can be useful for reputation system developers in order to preview and analyse new forms of attack, as well as to help to develop effective defence mechanisms.

Our proposal is still in its first version. Due to its extensibility, new attributes and values can be added to the framework. The reputation systems still have open issues, motivating us to continue our work. Thus, we expect to present new versions of this classification framework in the near future.

REFERENCES

Ahn, J., Park, M., & Paek, J. (2018). Reptor: A Model for Deriving Trust and Reputation on Blockchain-based Electronic Payment System. 2018 International Conference on Information and Communication Technology Convergence (ICTC), 1431–1436. https://doi.org/10.1109/ICTC.2018.8539641

Ahn, J., Park, M., Shin, H., & Paek, J. (2019). A Model for Deriving Trust and Reputation on Blockchain-Based e-Payment System. Applied Sciences, 9(24), 5362. https://doi.org/10.3390/app9245362

Alshammari, S. T., Albeshri, A., & Alsubhi, K. (2021). Building a trust model system to avoid cloud services reputation attacks. Egyptian Informatics Journal, 22(4), 493–503. https://doi.org/10.1016/j.eij.2021.04.001

Briner, R. B., Denyer, D., & Rousseau, D. M. (2009). Evidence-Based Management: Concept Cleanup Time? Academy of Management Perspectives, 23(4), 19–32. https://doi.org/10.5465/AMP.2009.45590138

Camilo, G. F., Rebello, G. A. F., de Souza, L. A. C., & Duarte, O. C. M. B. (2020). A Secure Personal-Data Trading System Based on Blockchain, Trust, and Reputation. 2020 IEEE International Conference on Blockchain (Blockchain), 379–384. https://doi.org/10.1109/Blockchain50366.2020.00055

Damiani, E., di Vimercati, D. C., Paraboschi, S., Samarati, P., & Violante, F. (2002). A reputation-based approach for choosing reliable resources in peer-to-peer networks. Proceedings of the 9th ACM Conference on Computer and Communications Security - CCS '02, 207. https://doi.org/10.1145/586110.586138

Dellarocas, C. (2000). Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. Proceedings of the 2nd ACM Conference on Electronic Commerce - EC '00, 150–157. https://doi.org/10.1145/352871.352889

Dellarocas., C. (2000). Mechanisms for coping with unfair ratings and discriminatory behavior in online reputation reporting systems. Proceedings of the Twenty First International Conference on Information Systems (ICIS '00), 520–525.

Dellarocas, C. N. (2003). The Digitization of Word-of-Mouth: Promise and Challenges of Online Feedback Mechanisms. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.393042

Dennis, R., & Owen, G. (2015). Rep on the block: A next generation reputation system based on the blockchain. 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST), 131–138. https://doi.org/10.1109/ICITST.2015.7412073

Dhakal, Anup, & Cui, Xiaohui. (2019). DTrust: A Decentralized Reputation System for E-commerce Marketplaces.

Dixon-Woods, M., & Fitzpatrick, R. (2001). Qualitative research in systematic reviews: Has established a place for itself. British Medical Journal, 323, 765–766.

Douceur, J. R. (2002). The Sybil Attack. In: Druschel, P., Kaashoek, F., Rowstron, A. (eds) Peer-to-Peer Systems. Lecture Notes in Computer Science, Vol 2429. Springer, Berlin, Heidelberg, 2429.

Feng, Q., Liu, L., & Dai, Y. (2012). Vulnerabilities and countermeasures in context-aware social rating services. ACM Transactions on Internet Technology, 11(3), 1–27. https://doi.org/10.1145/2078316.2078319

Fraga, D., Bankovic, Z., & Moya, J. M. (2012). A Taxonomy of Trust and Reputation System Attacks. 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, 41–50. https://doi.org/10.1109/TrustCom.2012.58

Gambetta, D. (1988) Can we Trust Trust? Gambetta, D., Ed., Trust: Making and Breaking Cooperative Relations. Blackwell, New York, 213-237

Gong, Y., van Engelenburg, S., & Janssen, M. (2021). A Reference Architecture for Blockchain-Based Crowdsourcing Platforms. Journal of Theoretical and Applied Electronic Commerce Research, 16(4), 937–958. https://doi.org/10.3390/jtaer16040053

Hendrikx, F., Bubendorfer, K., & Chard, R. (2015). Reputation systems: A survey and taxonomy. Journal of Parallel and Distributed Computing, 75, 184–197. https://doi.org/10.1016/j.jpdc.2014.08.004

Hoffman, K., Zage, D., & Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. ACM Computing Surveys, 42(1), 1–31. https://doi.org/10.1145/1592451.1592452

Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. Decision Support Systems, 43(2), 618–644. https://doi.org/10.1016/j.dss.2005.05.019

Karode, T., Werapun, W., & Arpornthip, T. (2020). Blockchain-based Global Travel Review Framework. International Journal of Advanced Computer Science and Applications, 11(8). https://doi.org/10.14569/IJACSA.2020.0110813

Koutrouli, E., & Tsalgatidou, A. (2012). Taxonomy of attacks and defense mechanisms in P2P reputation systems—Lessons for reputation system designers. Computer Science Review, 6(2–3), 47–70. https://doi.org/10.1016/j.cosrev.2012.01.002

Koutrouli, E., & Tsalgatidou, A. (2016). Reputation Systems Evaluation Survey. ACM Computing Surveys, 48(3), 1–28. https://doi.org/10.1145/2835373

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Reprint—Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Physical Therapy, 89(9), 873–880. https://doi.org/10.1093/ptj/89.9.873

Panagopoulos, A., Koutrouli, E., & Tsalgatidou, A. (2017). Modeling and Evaluating a Robust Feedback-Based Reputation System for E-Commerce Platforms. ACM Transactions on the Web, 11(3), 1–55. https://doi.org/10.1145/3057265

Petticrew, M., & Roberts, H. (. (2005). Systematic reviews in the social sciences: A practical guide (1 edition (M. A. Malden, Ed.; 1st ed.). Oxford: Wiley-Blackwell.

Sänger, J., Richthammer, C., & Pernul, G. (2015). Reusable components for online reputation systems. Journal of Trust Management, 2(1), 5. https://doi.org/10.1186/s40493-015-0015-3

Schaub, A., Bazin, R., Hasan, O., & Brunie, L. (2016). A Trustless Privacy-Preserving Reputation System (pp. 398–411). https://doi.org/10.1007/978-3-319-33630-5_27

Swamynathan, G., Almeroth, K. C., Ben, ·, Zhao, Y., Swamynathan, G., Almeroth, · K C, & Zhao, B. Y. (2010). The design of a reliable reputation system. Springer, 10(3), 239–270. https://doi.org/10.1007/s10660-010-9064-y

Thomas, J., Gough, D., & Oliver, S. (2017). Introduction to Systematic Reviews (2nd ed.). SAGE Publications, Limited.

Wang, J., Jing, X., Yan, Z., Fu, Y., Pedrycz, W., & Yang, L. T. (2020). A Survey on Trust Evaluation Based on Machine Learning. ACM Computing Surveys (CSUR), 53(5). https://doi.org/10.1145/3408292

Yao, Y., Ruohomaa, S., & Xu, F. (2012). Addressing Common Vulnerabilities of Reputation Systems for Electronic Commerce. Journal of Theoretical and Applied Electronic Commerce Research, 7(1), 3–4. https://doi.org/10.4067/S0718-18762012000100002

Zeynalvand, L., Luo, T., Andrejczuk, E., Niyato, D., Teo, S. G., & Zhang, J. (2021). A Blockchain-Enabled Quantitative Approach to Trust and Reputation Management with Sparse Evidence. Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '21).

Zulfiqar, M., Tariq, F., Janjua, M. U., Mian, A. N., Qayyum, A., Qadir, J., Sher, F., & Hassan, M. (2021). EthReview: An Ethereum-based Product Review System for Mitigating Rating Frauds. Computers & Security, 100, 102094. https://doi.org/10.1016/j.cose.2020.102094