Survey

Taxonomy of attacks and defense mechanisms in P2P reputation systems—Lessons for reputation system designers

Eleni Koutrouli*, Aphrodite Tsalgatidou

Department of Informatics and Telecommunications, National & Kapodistrian University of Athens, Panepistimiopolis 157 71 Ilisia, Athens, Greece

Article Info

Article history:
Received 23 September 2010
Received in revised form
24 January 2012
Accepted 31 January 2012

Keywords:
Reputation systems
Reputation attacks
Threat analysis
Credibility
Reputation attacks taxonomy
Peer-to-Peer

Abstract

Robust and credible reputation systems are essential for the functionality of Peer-to-Peer (P2P) applications. However, they themselves are susceptible to various types of attacks. Since most current efforts lack an exploration of a comprehensive adversary model, we try to fill in this gap by providing a thorough view of the various credibility threats against a decentralized reputation system and the respective defense mechanisms. Therefore, we explore and classify the types of potential attacks against reputation systems for P2P applications. We also study and classify the defense mechanisms which have been proposed for each type of attack and identify conflicts between defense mechanisms and/or desirable characteristics of credible reputation systems. We finally propose a roadmap for reputation system designers on how to use the results of our survey for the design of robust reputation systems for P2P applications.

© 2012 Elsevier Inc. All rights reserved.

Contents

1. Introduction ................................................................................................................................................................................. 48
2. Concepts and applications of P2P reputation systems .................................................................................................................. 50
  2.1. Basic concepts of P2P reputation systems .......................................................................................................................... 50
  2.2. Credibility of reputation systems ........................................................................................................................................ 52
    2.2.1. Factors related to the quality of recommendation creation and recommendation content ............................................ 52
    2.2.2. Factors related to the quality of recommendation selection .......................................................................................... 52
    2.2.3. Factors related to the quality of reputation reasoning ................................................................................................. 53
3. Taxonomy of attacks .................................................................................................................................................................... 53
  3.1. Unfair recommendations ................................................................................................................................................... 54
    3.1.1. Unfair recommendations from individuals ...................................................................................................................... 54
    3.1.2. Collusion .............................................................................................................................................................. 54
1574-0137/$ - see front matter © 2012 Elsevier Inc. All rights reserved.

* Corresponding author.
E-mail addresses: ekou@di.uoa.gr, ekoutrouli@bankofgreece.gr (E. Koutrouli), atsalga@di.uoa.gr (A. Tsalgatidou).
1. Introduction

In the last few years, a lot of research activity has been targeting Peer-to-Peer (P2P) technology, systems, architectures and applications. Such systems comprise a number of autonomous heterogeneous peers with intermittent presence in the network and a high level of anonymity, which interoperate in applications such as file sharing (e.g. [1–4]), instant messaging (e.g. [5]), distributed computing (e.g. Seti@Home [6]), or P2P e-commerce (e.g. [7,8]). These applications exploit the basic characteristics of P2P technology, namely scalability, self-organization, high level of distribution and ability to function in the presence of a highly transient population of nodes, network, and computer failures, without the need of a central server and central administration. This decentralized nature and the lack of a controlling authority expose P2P systems\(^1\) to a broad range of security attacks, as it has been

\(^1\) In the rest of the paper we use the terms 'P2P systems' and 'P2P applications' interchangeably.
widely recognized by researchers [9–11]. Examples of such attacks include:

- Denial of Service (DoS), where adversaries consume or cause waste of network-level resources (bandwidth, buffer space) and application-level resources (computation, memory) in order to prevent clients of the system from obtaining timely service.

- Free-loading, where adversaries attempt to benefit from the system’s services while contributing as less as possible to the system.

- Poisoning and pollution attacks, where adversaries inject corrupted or mislabeled content into a P2P file sharing application deliberately or accidentally respectively, resulting in decreasing the availability of shared items [12,13].

- Other malicious actions, where attackers attempt to compromise the integrity of the content or of the service provided by a P2P system or the access to restricted resources by not complying to the P2P routing or storage protocols, as shown in [9]. For example, in Eclipse Attack [14] a set of malicious peers arrange for a node of an overlay P2P network to be connected only with malicious peers and thus cut the peer off the network.

The effect of these attacks is maximized when malicious peers cooperate with each other or when a peer enters a P2P system with multiple identities (Sybil attack [15]) and orchestrates them in order to subvert the system.

The aforementioned attacks cannot be alleviated by traditional centralized security mechanisms which usually protect resources by assigning rights to entities based on their identities. Therefore, a number of advanced decentralized security mechanisms have been proposed in order to protect a P2P system from various attacks; examples include mechanisms for secure routing and secure message forwarding [9], mechanisms for DoS defense [11] and so on.

In addition, peers need to make trust decisions about whom to transact with, i.e. which peer’s services they can rely on. Trust decisions are also needed for selecting honest peers to participate in P2P functions, such as routing and message forwarding, in order to avoid some of the aforementioned attacks. In P2P content sharing applications trust decisions may also be needed about which object to choose among the objects offered, so as to choose an authentic one and to avoid pollution and poisoning (e.g. [12,16]). Therefore, trust is a prominent issue in P2P computing, and necessitates the use of trust mechanisms. Reputation systems, also referred to as reputation-based trust systems, have emerged to address this need and constitute an important trust management mechanism in online communities. Their main aim is to support trust decisions, i.e. to help an entity to decide whether to trust another entity and have a transaction with it or not, by using past behavior as a predictor of future behavior. However, reputation systems suffer themselves from attacks coming mainly from unfair or deceitful raters, which, for example, could exploit loose registration and authentication policies in order to deceive the system without being identified. Such attacks can be quite synchronized and powerful and reduce the credibility of the reputation system.

Several reputation systems exist in the research literature and a number of surveys related to them have been presented, as in [17–21]. However, work focusing on the credibility of reputation systems and their resistance to comprehensive adversary models is still in early stages. The work in [22], which presents a threat-centric framework for evaluating and comparing a number of decentralized reputation systems, and the work in [23] which presents an analysis and comparison of reputation systems regarding basic credibility criteria are two of the few examples found in this area. These works however focus on specific attacks and credibility issues and do not thoroughly investigate the threats against reputation systems and the relative defense mechanisms. Baker et al. [24] and the more recent work by Hoffman et al. [25] present a security analysis for reputation systems and a survey of attacks and defense mechanisms in reputation systems for P2P applications respectively. More specifically, Baker et al. [24] present the main characteristics and various use cases of electronic reputation systems, the main threats against them and a set of recommendations for best practices in their use. In the survey of Hoffman et al. [25], the authors identify the main attacks against reputation systems for P2P applications and decompose reputation systems to their general components for identifying which particular components are the target of each attack and then present existing defense mechanisms. The first of the aforementioned surveys tackles attacks against reputation systems in general, while the second one identifies a set of broad attacks against reputation systems for P2P applications without elaborating on these attacks. The review of the literature shows that the work done so far in the field of attacks against reputation-based trust systems for P2P applications is not comprehensive enough and lacks a systematic analysis of the attack space related to reputation-based trust systems for P2P applications.

The main goal of this paper is to fill in the aforementioned gap by: (a) thoroughly investigating and appropriately classifying the types of attacks against reputation-based trust systems for P2P applications in order to better understand the weaknesses of such reputation systems, and (b) exploring the ways that can be used to enhance the robustness of reputation systems against those attacks. Thus, the rest of the paper presents the results of our work towards this goal. These results have been derived through a thorough examination of a variety of reputation systems for P2P applications (P2P reputation systems) with respect to both their susceptibility and robustness against specific threats. The sections that follow have been organized as follows: Section 2 presents the basic concepts of P2P reputation systems and the types of applications that can benefit from those systems; the concept of reputation system credibility and the associated requirements and credibility factors are also analyzed. Section 3 examines and classifies the various attacks towards reputation systems. Section 4 presents the existing defense mechanisms against those attacks. In Section 5 we analyze the proposed taxonomy, discuss about the conflicts and tradeoffs between desirable features of reputation systems and propose guidelines for reputation system designers on how to build attacks resistant reputation systems for P2P applications; conclusions follow in Section 6.
2. Concepts and applications of P2P reputation systems

A rich variety of reputation systems has been proposed in the literature based on either a centralized or a decentralized architecture. Centralized reputation systems (e.g. eBay [26]) rely on a central control or administration of information, whereas in decentralized reputation systems (e.g. [27–30]) reputation information management is distributed to all participating entities. In the latter case, some entities could still have a special role, acting for example as trusted parties. Decentralized reputation systems are more suitable for P2P applications where no central entities exist. In our survey of threats and defense mechanisms against reputation systems we have focused on decentralized reputation systems; however, in order to make our analysis more comprehensive, we have also taken into consideration the literature regarding the centralized reputation system of eBay [26] and some approaches which assume some form of central or semi-central infrastructure (e.g. [30–32]) when the results are applicable or extensible to the decentralized case.

2.1. Basic concepts of P2P reputation systems

In a decentralized reputation system the participating entities play interchangeably the roles of trustor, trustee and recommender. The trustor is an entity which wants to make a trust decision regarding whether to participate or not in a transaction with another entity, the trustee. A transaction can involve accessing or allowing access to a resource, e.g. a file, buying or selling goods, etc. The recommender is the entity that provides the trustor with information regarding the trustworthiness of the trustee; this information is known as recommendation. In file sharing P2P applications, recommendations may also be given for objects, e.g. files; reputation of objects may be estimated to help trust decisions about which object is authentic and thus can be chosen when a particular object is required.

To make a trust decision the trustor tries to predict the future behavior of the trustee by forming a view of the trustee based on experience about its earlier actions. This subjective view is formed by estimating an indicator of the quality of the trustee regarding its services (e.g. provision of a file, an ecommerce trade) and comprises the trustee’s reputation or trustworthiness from the trustor’s point of view. To form a reputation view, the trustor needs to gather experience information, either by referring to its own earlier experience with the trustee, or by acquiring it from other entities in the form of recommendations. Recommendations can be based on the recommender’s personal experience alone, or on a combination of personal experience and earlier recommendations from others. A recommendation is either a rating describing a single transaction experience (transaction-based recommendation), or an opinion formed by the outcome of several transactions and possible outside experience (opinion-based recommendation). The aggregation of recommendations regarding the trustee results in estimating the trustee’s reputation value. Apart from explicit feedback in the form of recommendations, reputation estimation could also take into consideration other information by analyzing attributes of the ‘capability’ of a peer, such as bandwidth, processing power, memory and storage capacity [33]. Reputation values can then be translated in a way that facilitates a trust decision.

A trust decision is either threshold-based (e.g. [27,29,34]) where the selection of the trustee for a transaction depends on the comparison between a specific threshold and the trustee’s reputation value, or rank-based, where the trustor compares the trustworthiness of different peers and chooses to transact with the entity which has the highest reputation value (e.g. [35]). A slightly different rank-based trust decision method is proposed by Papaioannou and Stamoulis [36]; in this approach, a peer (service requester) selects another peer (service provider) to transact with using a probabilistically fair manner according to the rank of the latter peer. A peer’s rank is estimated based on its own reputation and the average reputation of all peers and determines the demand that this peer can probabilistically attract as a future service provider.

The reputation of a peer is context-specific. Context may refer to the kind of service the trustee provides (e.g. a peer could be highly reputable when coming to providing files but have a low reputation as a recommender provider) or with regards to specific attributes of such a service (authenticity of provided files, quality and quantity of a product purchased by the trustee in an e-commerce trade, etc.). Reputation also depends on time, as the behavior of peers is dynamic; furthermore, recent transactions are often considered more important for reputation estimation than older ones.

A reputation system may support one or multiple levels of feedback transitivity. In the case of a single transitivity level, a trustor directly collects information from peers which have had transactions with the trustee. In the case of multiple transitivity levels, the trustor indirectly collects information from peers which have had transactions with the trustee. This means that chains of entities are formed to communicate to the trustee recommendations from entities which have directly interacted with the trustee. In such chains of entities, each pair of two successive entities have transacted with each other, so every peer has estimated a reputation value for the previous peer and all these values are used for the estimation of the trustee’s reputation value. Reputation systems with multiple transitivity levels, are often graphically presented as directed graphs, as in [37], where nodes represent the peers, edges represent transactions between the connected peers and the weight of an edge represents the context of the corresponding transaction and the rating of the behavior of the destination node in the transaction as perceived by the source.

Reputation values are estimated either locally and subjectively by the trustor (e.g. [27,38]) or as global values calculated by special peers (e.g. [30,39,40]). In the first case every peer estimates personalized reputation values for other peers, based on its own opinions and selected recommendations from third parties. In this way, one peer has different reputation in different peers’ database, as it happens in real life interactions. In the latter case, a unique global reputation value is estimated for each peer in the network, based on the opinions from the whole peer population. Local reputation values are stored by the trustor itself, whereas for global reputation values ‘storing peers’, i.e. peers storing the reputation values, are determined either randomly [41] or by using various techniques, such as Bloom
filters, as in [39] or a Distributed Hash Table (DHT),\textsuperscript{2} as in [30]. A different approach for reputation estimation and reputation information storage is followed in NICE [44], where the trustee is the entity which estimates its own reputation value from a chain of entities which have interacted with each other and connect the trustor with the trustee. The trustee then stores its own reputation value.

The various roles of the participating entities in a decentralized reputation-based trust system are illustrated in the UML diagram of Fig. 1.

Based on the above and on the results of some further analysis that we conducted in the field, we outline below the basic characteristics of a reputation system and how they can be instantiated:

(a) Recommendation content and its representation. This could be an arithmetic value, a statement, or a combination of a value and associated semantic information, such as the confidence of the recommender, context and time information, etc. A recommendation can represent only positive behavior, only negative behavior or both types of behavior of the trustee. Regarding its representation, various formats can be used, such as binary, scalar or continuous values in a specific interval, such as [0, 1], or text in case that it contains semantic information (as in [45]);

(b) Recommendation formation. A recommendation can be formed based on the evaluation of a single transaction that the recommender conducted with the trustee or on aggregated ratings regarding the transactional behavior of the trustee. Techniques that can be used for the formation can vary from manual ratings of single transactions (as in [27]) to probabilistic techniques (as in [40]) and statistical aggregation of simple arithmetic ratings (as in [30]).

(c) Selection of recommenders. Recommenders can be selected based on their credibility as recommenders, on knowledge regarding the social relationships between peers, on the similarity of the recommenders’ previous recommendations with the recommendations of the trustor regarding commonly evaluated peers, and so on.

(d) Reputation estimation. It refers to the calculation approach used for transforming the available information (both direct transactional information and recommendations from others) to a useful reputation metric. As described in [46], such an approach can be either deterministic, probabilistic or based on fuzzy logic.

(e) Storage and dissemination of reputation information. Reputation values may be estimated either reactively (i.e. on demand) or proactively (i.e. after each transaction conducted by the trustee). Calculated reputation values are stored either locally by the trustor or the trustee or by other peers (which we call ‘storing peers’) that are determined using

\textsuperscript{2} Examples of DHTs include [42,43].
various techniques as we have aforementioned. A reputation value may be stored by more than one ‘storing peers’ for preserving integrity. Furthermore, reputation values are communicated by the storing peers to other interested parties either upon request, or using a disseminating technique, such as flooding [44].

The way a trust decision is made. Trust decisions are threshold-based or rank-based; they are based on the estimated reputation values, therefore, the latter should be translated in a manner that facilitates trust decisions.

Various reputation systems have been proposed for different categories of P2P applications, such as P2P communities [27, 44], P2P e-commerce [28, 29, 34] and file sharing [12, 16]. Each application category exhibits specific context characteristics which affect the choices of a reputation system designer regarding the aforementioned characteristics of a reputation system (e.g. content and formation of the recommendation or the way the recommender is selected). A comparison of the various design choices in a number of reputation systems for different P2P application types is presented in [47]. Furthermore, as proposed by Dellarocas [48], design choices for reputation systems depend also on the settings of the application environment. Dellarocas makes a distinction between the case when transactions are made between self-interested entities which may strategically choose the quality of service they provide and the case when the behavior of an entity is determined by its attributes (e.g. its bandwidth or its intrinsic honesty). Reputation systems have a different role in each case; in the first case their role is to promote cooperative and honest behavior (sanctioning reputation systems), while in the second case their role is to promote learning about attributes of the entities, such as their capability or honesty (signaling reputation systems). Depending on the role of the reputation system the reputation system designer should make the suitable decisions regarding the aforementioned reputation system characteristics.

2.2. Credibility of reputation systems

Credibility is an essential property of a reputation system, and refers to the confidence that can be placed on its effectiveness. An effective reputation system is expected to estimate the peers’ trustworthiness as accurately as possible in order to lead to the right trust decisions. This estimation, and thus the credibility of a reputation system is affected by a number of factors which are analyzed in the following paragraphs and have been grouped in the following three categories:

- Factors related to the quality of recommendation creation and recommendation content.
- Factors related to the quality of recommendation selection.
- Factors related to reputation reasoning.

Before analyzing these factors, we would like to mention that the credibility requirements of a reputation system depend on the type of the supported application, as the level of inherent risk level involved in a transaction is different in each application category; for example, the risk involved in a P2P e-commerce application transaction is higher than the one involved in a P2P file sharing application. The higher the transactional risk, the more attention that should be paid to the credibility issues of a reputation system.

2.2.1. Factors related to the quality of recommendation creation and recommendation content

Recommendation information should be carefully created so as to reflect the quality of real transactions with the trustee. In the case of opinion-based recommendations, where individual ratings are aggregated to form an opinion, both the aggregation method and the number of aggregated recommendations influence the credibility of the latter. For example, the more the ratings which are aggregated to produce an opinion are, the more credible the recommendation is. Furthermore, an aggregation method may take into consideration the time that transaction took place, in order to evaluate individual ratings based on their recency. Such an aggregation method produces a more accurate opinion-based recommendation, since recent transactions are most important as we have aforementioned.

Recommendation information can be of various types, for example binary, continuous or discrete and can reflect either positive or negative experiences or both, as described in [47]. As it will be examined in the following sections, the kind of information included in a recommendation (e.g. level of satisfaction or dissatisfaction) favors some types of attacks; for example when only negative experiences are recorded, then positive recommendations which are dishonest do not affect the credibility of the reputation system, whereas the system is vulnerable to badmouthing attacks.

Recommendations should also be well documented, ensuring ascertainable data provenance; for example they should provide the ability to trace the source of a recommendation and the originality of transactions in case of transaction-based recommendations. Actually, recommender traceability is important both in the case of a single feedback transitivity level and in the case of multiple transitivity levels. Time is also a useful documentation characteristic, as it was mentioned above; it could prove that a transaction-based recommendation refers to a real transaction; it could also be used to weight recommendations according to their recency. Recommendation information should additionally include some other metadata information, such as the confidence a recommender has for its opinion-based recommendation, depending on the amount of information it has.

2.2.2. Factors related to the quality of recommendation selection

The quality of recommendations depends highly on the quality of the recommenders who need to be the most relative and credible; therefore, attention needs to be paid to the method used for selecting recommenders (e.g. by taking into account knowledge about the social relationships between peers or by using recommendations from all recommenders), possible bias in this method and keeping track of the credibility of the recommender. Considerations about uncertainty and about possible forms of pressure that could bias recommendations should also be taken. For example, reciprocity could lead to insincere recommendations, as shown by Dellarocas et al. [49] and by Masclet and Pénard [50] for the case of eBay. When recommendation chains are used between the source of recommendation and the trustor, that is there are mediators which communicate recommendations to the trustor, the mediator’s credibility is also relevant. Storing and retrieving recommendation information should also be done in a secure manner, which
The quality of reasoning of a reputation system is affected by:

- the algorithm used for the aggregation of recommendations, in order to estimate a reputation value, e.g., aggregation formula and whether and how recency and context of recommendations is taken into account;
- the way the calculated reputation measure is translated in order to enable a trust decision, which, as we have mentioned before, can be a rank-based or a threshold-based decision. In the case of threshold-based trust decision, selecting a suitable threshold is also important;
- the kind and amount of information history stored and used; and
- the evaluation of a calculated reputation value, so as to reflect the confidence that can be placed on it. The level of confidence related to a reputation value may depend on the amount, the quality and the variability of aggregated recommendations.

Furthermore, when a reputation system calculates global reputation values, storing and retrieving these values should be done in a secure manner, so that information is not lost, altered or intercepted.

The potential types of attacks against a reputation system depend on the design decisions regarding the aforementioned credibility factors. Table 1 summarizes these factors and organizes them into the three categories presented above.

### Table 1 – Credibility factors of a reputation system.

<table>
<thead>
<tr>
<th>Factors related to recommendation creation/content</th>
<th>Factors related to recommendation selection</th>
<th>Factors related to reputation reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Type of recommendation information (value, statement, etc.)</td>
<td>- Recommender’s credibility</td>
<td>- Aggregation method (estimation formula, recency considerations, reputation value translation)</td>
</tr>
<tr>
<td>- Creation method (transaction rating or opinion)</td>
<td>- Uncertainty awareness</td>
<td>- History of transactions and recommendation information</td>
</tr>
<tr>
<td>- Type of experience (negative and/or positive) evaluated in a recommendation</td>
<td>- Recommender selection method, considerations about possible bias or pressure</td>
<td>- Storage and dissemination methods for reputation values</td>
</tr>
<tr>
<td>- Time of recommendation/time of transaction</td>
<td>- Storage and dissemination methods for recommendations, considerations about possible bias or pressure</td>
<td>- Evaluation of estimated reputation</td>
</tr>
<tr>
<td>- Recommender’s identity</td>
<td>- Mediator’s credibility</td>
<td>- Secure storage and retrieval of global reputation values</td>
</tr>
<tr>
<td>- Recommender’s confidence on recommendation</td>
<td>- Who collects recommendations, possible bias</td>
<td></td>
</tr>
<tr>
<td>- Binding recommendations with transactions</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3. Taxonomy of attacks

Entities participating in reputation systems can distort the credibility of the latter in various ways, either deliberately or not, isolated or in cooperation with others, depending on the specific application and social setting of the reputation system. Reputation attacks or entities misbehavior can be classified in the following three main categories:

- **Unfair recommendations**: Entities can spread unfair ratings for other entities or can do it with cooperation with each other to maximize the effect of the attack. Unfair ratings can be due to lying, misjudging the outcome of a transaction, or making a mistake in the recommendation procedure.
- **Inconsistent behavior**: Peers may strategically have an inconsistent behavior that can lead to an incorrect estimation of their reputation allowing them to misbehave and still keep a high reputation. They can, for example misbehave part of the time or towards a subset of peers or change their behavior suddenly or periodically.
- **Identity management related attacks**: A deciding factor for attacks in this category is the identity scheme used in a reputation system. For example, when the identity scheme permits the use of multiple identities by the same peer, a malicious peer can have a dishonest behavior and then escape its low reputation by entering the system with a new identity. Furthermore, when an entity A can communicate or store a recommendation produced by an entity B for an entity C without linking its identity and B’s identity with the recommendation, then A can easily manipulate the rating by changing or omitting the recommendation value. Also, if the system permits it, peers can refuse having sent a rating, which makes them capable of sending unfair ratings without taking the responsibility of their action, making thus the reputation system unreliable.

Sometimes, an attack may be a combination of the above categories in order to cause a more serious damage. In the rest of this section we describe these categories in detail and divide them in further subcategories. It is worth to mention here that apart from the aforementioned categories of attacks which target the reputation system itself, a P2P reputation system is vulnerable to the attacks that have been discussed...
in Section 1 which target P2P systems themselves and which have been extensively studied by the literature [9–11]. In our paper we examine the attacks which particularly target P2P reputation systems and the respective defense mechanisms.

3.1. **Unfair recommendations**

Recommendations which do not reflect the true quality of a transaction or an honest opinion about the transactional behavior of the trustee are characterized as unfair recommendations. They can be due to the recommender being either malicious or self-interested or having incomplete information. Depending on when unfair recommendations are sent by individual peers or by strategically acting collusions of peers, we have organized the various types of attacks belonging to this category as follows:

3.1.1. **Unfair recommendations from individuals**

Single peers may send unfair negative or positive recommendations (‘bad mouthing’ or ‘unfair praises’ accordingly), random opinions or inaccurate recommendations, as described in the following types of attacks:

3.1.1.1. **Bad mouthing.** A malicious peer can ‘bad-mouth’ other peers (i.e. spread fraudulently low recommendations for them) in order to unfairly reduce their reputation, so as to increase its own reputation when related to them. A specific type of bad mouthing consists in *discriminating* when providing recommendations. In this case, a malicious entity provides fair recommendations except when dealing with a specific partner. Thus, it manages to be regarded as a good recommender and at the same time to harm effectively its target.

3.1.1.2. **Random opinions.** In the case of reputation systems which reward peers for the provision of recommendations (e.g. [51]), selfish peers could send random opinions which do not correspond to genuine transactions in order to gain the rewards. Peers may also prefer to give random opinions about other peers instead of honestly evaluating them, as an honest evaluation of a participant’s performance costs more in time and resources than a random one. Furthermore, honest (and positive) recommendations by a high reputable recommender could give the recommended peers an advantage over the recommender, as the reputation of the recommended peer is being increased. It is thus clear that, if both random and honest recommendations result to the same reward, not only there is no incentive for active participation but there is an incentive for inaccurate feedback.

A large number of random opinions could also be sent by malicious peers aiming at causing a *Denial of Service (DoS)* attack, as a particularly large number of recommendations should then be distributed and aggregated by the reputation system, increasing thus the communication cost and degrading its performance.

3.1.1.3. **Inaccurate recommendations.** These can be due to having incomplete information, e.g. a peer which sends an opinion-based recommendation about another peer may have little experience with it, and, thus a *weak confidence* about the opinion. Such a recommendation cannot be considered as accurate as it involves a high level of uncertainty. In case that the reputation system supports the provision of the recommender’s confidence, this attack can be handled during reputation estimation by for example ignoring or giving less importance to recommendations with weak confidence. Inaccurate recommendations can also be due to sending a wrong rating or opinion by mistake.

3.1.1.4. **Unfair praises under pressure or strategic considerations.** Participants of reputation systems may give higher recommendations than what the trustee deserves, under fear of reprisals or reciprocity expectation. Dellarocas et al. [49] provide empirical evidence of these phenomena on eBay aiming at determining the drivers of trader participation in the eBay reputation system. This evidence shows that the result of the fear of reprisals and the expectation for reciprocal behavior is the artificial increase of the number of positive evaluations. As an example, a buyer or a seller can submit an unjustified positive evaluation, for the purpose of encouraging the transaction partner to send back a positive evaluation. Similarly, a participant may not give a *justified* negative evaluation out of fear of receiving an ‘unjustified’ negative evaluation in return.

3.1.2. **Collusion**

This category of attacks refers to the threats posed when a group or groups of malicious peers try to subvert the system. The damage to the credibility of a reputation system caused when multiple malicious peers are acting together is in most cases more serious than the damage caused when each peer is acting malevolently independently. Specific examples of collusion attacks are the following:

3.1.2.1. **Collusive badmouthing.** Some malicious peers can cooperate with each other in order to spread negative ratings about an honest peer and seriously decrease its reputation. At the same time they can give positive ratings about each other and, thus, increase their own reputation.

3.1.2.2. **Collusive reducing recommendation reputation.** Coordinated malicious nodes can also reduce the recommendation reputation of honest peers by badmouthing only a subset of the entities they transact with. In this way, they create conflicting opinions about both the transactional behavior of the victims and of the recommendation reputation of the honest recommenders which recommend the victims. At the same time their own recommendation reputation is not affected seriously, as they spread honest recommendations for the rest of the entities.

3.1.2.3. **Collusive deceit.** All entities of a group of malicious peers behave badly, but provide positive recommendations for each other. In a variation of this attack, a single entity of the collusive group behaves dishonestly while the others spread positive recommendations for it. In a more complicated variation, a single entity misbehaves only some of the time in order to escape detection while the other members of the group spread positive recommendations for it. *Ballot stuffing* is a term frequently used to indicate an attack of this category. It refers to the situation when ratings for fake transactions are provided. For example, in an online auction
system a seller can collude with a group of buyers in order to be given unfairly high ratings, which do not correspond to real transactions. This will have the effect of inflating its reputation, allowing that seller to receive more orders from other buyers and sell at a higher price than deserved.

3.2. Inconsistent behavior

As we have aforementioned, peers can have an inconsistent behavior for obtaining positive reputation while acting in a strategic manner. Inconsistency refers either to the transactional behavior of the attacker or to the relation between its transactional and recommending behavior, as shown in the following categories:

3.2.1. Traits

Some peers may behave properly for a period of time in order to build up a strongly positive reputation, so as to be highly trusted, and then begin deceiving. Such an example is found in eBay, where sellers participate in many small transactions in order to build up a high positive reputation, and then cheat one or more buyers on a high-priced item. These sudden changes in the transactional behavior of a peer make it difficult for the other peers to reduce the attacker's reputation adequately. This category also includes oscillatory transactional behavior where a peer keeps changing its behavior from honest to dishonest; this makes it difficult to update the attackers' reputation timely and effectively.

3.2.2. Discrimination when providing services

A malicious entity may behave properly with most entities and misbehave towards one of them or a small subset of them. In this way, it can obtain a good reputation and still manage to harm its selective targets, without causing a considerable damage to its own reputation. For example, a seller can provide good quality to all buyers except one single buyer. Recommendations about that particular seller will indicate to an identity scheme, which, as stated before, should de-register to the system with some form of identity according to the identity policy followed and the anonymity level supported by a P2P system is also related with the level of anonymity supported, or paying a monetary registration fee, etc. As analytically described in the creation of an identity could involve some kind of cost, e.g. solving a resource consuming problem in order to register or paying a monetary registration fee, etc.

The identity management scheme followed by a P2P system is also related with the level of anonymity supported, and thus the level of privacy a peer could have, which is a desirable characteristic of most P2P systems. It is worth to note that the anonymity level supported by a P2P system is determined by the way that it handles issues like:

• identifying a peer which holds/ asks for receives a piece of trust information
• associating a particular transaction with a specific peer
• associating a specific pseudonym with a real-world identity.

It is obvious that the more information is linked to the identity of a peer (either a real-world identity or a pseudonym) the less the anonymity level of the peer and the privacy it has, and the higher the level of accountability and the credibility of the reputation estimation that can be achieved. Reputation is, thus, highly related in a trade-off manner with the level of privacy and in proportion to the level of accountability that can be achieved through the identity policies followed.

Attacks in this category exploit the type of identity management followed and the anonymity level supported by a system and can be classified as registration policy-related and authentication policy-related attacks.

3.3.1. Registration policy-related attacks

Each participating peer in a reputation system needs to register to the system with some form of identity according to an identity scheme, which, as stated before, should define the following characteristics:

• traceability (vs. anonymity) of a peer: the identity is either linked with the peer’s real-world identity or not (in this latter case it is known as pseudonym);
• method of identity creation: the identity can be created by the peer itself or assigned to the peer by the system, e.g. by using a Public Key Infrastructure (PKI) [52];
• cost association: creation of an identity can be associated to some form of cost or be free [53];
• uniqueness of a peer’s identity: a peer may have one unique identity or multiple identities, which can be either unlinked or linked to each other so as to refer to the same entity.

When a reputation system supports pseudonymity and there is no associated cost with the formation of new unlinked identities, the following types of attacks may occur:

3.3.1.1. Sybil attack. Sybil attack refers to a situation where a single malicious peer creates multiple identities and uses them in concert to defeat the system [15]. In this situation, a malicious peer can enter the system with different identities acting each time as a distinct peer and provide a large amount of false reputation information. As analytically described
in [54], Sybil-like attacks in a reputation system can target a single entity or a group of entities. In the first case, a malicious peer owning a collection of identities can use them to spread negative recommendations about a single user (self-collusion for bad mouthing) or deceive a single user by using some of its pseudonyms for ballot stuffing (self-collusion for ballot stuffing). It achieves that by dividing the identities into three groups: the service providers, the recommender which will give false positive opinions for the service providers and the ones which will increase the recommendation reputation of the recommends regarding their credibility. The victim selects the service provider based on its reputation and is deceived, while the attacker may not use the attacker’s pseudonym any more. When peers are organized in groups (e.g. groups of mutually trusted peers [32] or neighborhoods of peers [34]) and group reputation is taken into account for the estimation of a peer’s reputation, then a malicious peer can also harm the group reputation using the Sybil attack. For example, the attacker can join a group and then misbehave in order to lower the group’s reputation (insider attack), or can join different groups, build a good reputation as recommender and then spread negative recommendations about the target group (outsider attack), managing to lower its reputation.

3.3.1.2. Whitewashing/Pseudospoofer. A malicious peer can discard its identity and enter the system with a new one so as to escape bad reputations. A peer will have an incentive to do so, only if the reputation assigned to a new peer is greater than its own reputation. Such a peer is referred to as a whitewasher or as a pseudospoofer in case it changes its identity periodically to escape from being identified. In contrast to what is happening with the Sybil attack, pseudospoofer cannot easily be synchronized and coordinate to perform collusion attacks, as their distinct identities (created maliciously), cannot act at the same time. So, they can have one of the individual attacker behaviors, described in Section 3.1.1; however, they cannot conduct collusion attacks, even though they change their pseudonym periodically.

3.3.2. Authentication policy-related attacks
Further attacks can take place when no authentication of peers and messages is employed. Authentication of peers refers to the process of verifying the digital identity of the sender of a communication. When there is no way to ensure that the users of a reputation system are who they say they are and that the user who attempts to perform functions in a system is in fact the user who is authorized to do so, the following types of attacks can take place:

3.3.2.1. Impersonation. This attack refers to a malicious peer that portrays itself as another peer, stealing, for example, the victim’s pseudonym. In this way, the attacker can behave dishonestly on behalf of the impersonated peer (affecting, thus, the impersonated peer’s reputation negatively) or spread unfair ratings about others using the stolen pseudonym, (affecting thus, negatively the recommendation reputation of the impersonated peer).

3.3.2.2. Man-in-the-middle attack. In P2P systems peers need to rely upon intermediate peers to forward their queries or responses. This is the case in P2P reputation systems, where intermediate peers may have the opportunity to tamper with the responses, by for example:
- miscommunicating trust information (e.g. omitting or altering recommendations), or
- privacy breaching, i.e. using reputation information passing through them in order to infer the habits of the recommender.
This type of misbehavior is related to the inherent tradeoff between trust and privacy which is further explained in Section 5.

3.3.2.3. Repudiation. An entity may refuse that it has sent, asked for or received a recommendation. By not being able to verify that an entity is responsible for such actions, malicious peers have no fear of being identified and being punished and may choose to:
- issue unfair ratings;
- refuse to send a recommendation when they are asked to do so;
- send a great number of fake recommendations which can cause a congestion and performance degradation in the reputation system (Denial of Service attack); or
- unjustly accuse another entity of misbehaving (false accusation) during recommendation provision. For example, when a protocol for exchanging recommendations requires that an entity pays a fee when it receives a recommendation, the receiving entity can unjustly deny having either requested or received the recommendation and may thus refuse to pay [54]. In this case, the recommender is not rewarded for its recommendations and more importantly, its recommendation reputation is falsely decreased.

Fig. 2 summarizes the presented reputation attacks in a visual taxonomy, which is shown as a tree structure. The hierarchy of the nodes represents the hierarchy of the various categories of attacks, whereas the leaves represent the specific types of reputation attacks. This visualization facilitates the better understanding of the various cases of reputation attacks and can thus help a reputation system designer to acquire a good picture of the possible threats so that s/he can then focus on how to defend specific threats. The existing defense mechanisms against those attacks comprise the subject matter of the following section.

4. Defense mechanisms
In order to cope with the types of attacks presented in the previous section, a number of defense mechanisms have been proposed in the literature. These mechanisms aim at enhancing the robustness and credibility of reputation systems either by defending a particular type of attack or a combination of attacks. In this section we present these mechanisms and we classify them according to the specific attacks they address.

4.1. Defense mechanisms against unfair recommendations
A peer participating in a P2P reputation system should be able to assess the credibility of a recommendation passed on to
it, in order to determine whether to use it 'as is', ignore it or adjust it accordingly. We classify the defense mechanisms against unfair recommendations in the following three categories:

- Internal information-based methods.
- External information-based methods and
- Incentive-based defenses.

The defense mechanisms belonging to the first two categories just try to identify and isolate the unfair recommenders based on the recommendations and other information. They are referred to as endogenous and exogenous methods respectively in [55]. Mechanisms of the third category include reaction towards the recommending behavior, in order to give incentives for honest recommendations. This reaction can be punishment or rewards for dishonest recommenders and honest recommenders respectively. In the following we present the various defense mechanisms in each of these three categories.

4.1.1. Internal information-based methods
These methods rely on the provided recommendations in order to identify or prevent unreliable reputation information. They achieve this either by considering statistical properties of the recommendations, or by applying rules on the numbers of the aggregated recommendations. These methods include the following:

4.1.1.1. Similarity-based filtering techniques. Techniques in this category filter out recommenders whose recommendations have low similarity with the trustor's recommendations on commonly evaluated peers. For example, Dellarocas [31] uses collaborative filtering techniques to identify trusted recommenders based on their similarity with the trustor on commonly evaluated peers, considering recommenders whose opinions deviate from the trustor's opinion to be unreliable (from the trustor's point of view). This step may filter out the unfair recommenders except in the cases of ballot stuffing and positive discrimination, where the colluders have taken collaborative filtering into account and have sent recommendations which are similar to the trustor's recommendations for every other peer except the trustee; their recommendations for the trustee are significantly higher comparing with those of the trustor. A divisive clustering algorithm is, thus, subsequently used to separate the identified set of recommenders in two clusters based on their average ratings regarding the trustee, in order to exclude the cluster of agents which could have colluded and have given the trustee unfairly positive ratings. The work of Whitby et al. [56] falls in this category of defense mechanisms too. The authors compare long-run average recommendations and reject recommenders who have rated significantly different from the average over the time. Similarly, in Havelaar [57], recommendations for a peer which deviate from average recommendations for the same peer, are ignored when estimating the peer's reputation.

4.1.1.2. Aggregating recommendations from a large number of peers. This mechanism is used, when estimating the reputation of a peer, in order to downgrade the effect of bad mouthing. Furthermore, in this way, the damage that can be done by collusive badmouthing or collusive deceit is limited when the fraction of colluding peers is small [57].

4.1.2. External information-based methods
These methods rely on other information besides recommendations themselves, in order to determine the reliability of the recommendation. Such information may be the reputation of the recommender, or the relationship between the recommender
and the trustee. The following mechanisms are included in this category:

4.1.2.1. Estimating the reputation of the recommender (recommendation reputation). And using it as a weight for its recommendations or in order to decide whether to use the recommendation or not. Estimating the reputation values of mediators in recommendation chains and using them for reputation estimation is a subcase of these mechanisms which concerns reputation systems with multiple transitivity. Methods for recommendation reputation estimation are based either:

- on the reputation of the peer regarding its honesty in the transactions with other peers; or
- on the trustworthiness of a recommender regarding the recommendations it provides (recommendation reputation), rather than its trustworthiness regarding its provided services (service reputation).

Examples of methods in the first category are presented in PeerTrust [27], FuzzyTrust [28], EigenTrust [30] and NICE [44]. In FuzzyTrust [28], the reliability of a recommender is estimated as a function of its trustworthiness regarding the services it provides, while PeerTrust [27] and EigenTrust [30] use a peer’s trustworthiness as a weight for the recommendations it gives when estimating another peer’s trustworthiness. Similarly, in NICE the authors assume that the most trustworthy peers regarding their transactional behavior provide the most reliable recommendations.

Methods belonging to the second category of estimating the recommendation reputation are more sophisticated and vary in the way recommendation reputation is estimated and used. In general, these methods reduce the recommendation reputation of the entities which have recommended a dishonest entity, so as they will not be chosen as recommenders in the future or their recommendations will have little weight. In the rest of the section we present methods of this category. In [16] records are kept for the recommendation trustworthiness of a peer; these records are updated only in case the reputation of the trustee is updated after a transaction. In the work of Dillon et al. [29] each peer estimates the recommender’s trustworthiness regarding the recommendations it gives, as the difference between the given recommendation and the recommender’s value of its transactional behavior. In this way, in [58] the semantic distance between a peer’s recommendation and the recommender’s perception of the outcome of a transaction is proposed to adjust the recommendation. Zhao and Li propose H-Trust [59] where peers keep credibility ratings for other peers and update them based on the received recommendation and the result of the transaction with the recommended entity. The recommendations that a peer will use for calculating a trustee’s reputation are the ones that come from the most credible recommenders, whereas the other recommendations are ignored. In [60] credibility ratings are calculated by the trustee for each recommendation after the subsequent transactions between the trustee and the recommender. These credibility ratings are used to calculate the recommendation reputation of the recommender which will be used to weigh future recommendations from the specific recommender in the reputation calculation process.

In PeerTrust [27] recommendations are weighed by a credibility measure of the recommender for which one possible estimation method is based on the similarity between the recommender’s and the trustee’s opinions on commonly evaluated peers. Another example is the use of the perceived accuracy of the recommender’s past opinions, as in [55], where the probability that an agent will provide an accurate opinion is calculated given its past opinions, and subsequent observed interactions with the trustees. The accuracy of the recommendations of a particular entity is also used in [61], as a meta-trust value which is associated with each entity and weights recommendations provided by this entity, although its calculation mechanism is not provided in this work. A different approach is proposed by Jin et al. [62], where two transacting entities rate each other’s behavior in their transaction; if their ratings agree then they are considered consistent. A feedback reputation metric is thus calculated for each peer as the percentage of the number of consistent ratings the peer has given. After each transaction, the transactional reputation of a peer is updated based on whether the ratings of the two transacting entities agree and also on the feedback reputation of the other transacting entity.

4.1.2.2. Methods for unbiased selection of recommenders/recommendations. These methods may be based on social relationships, pre-trusted recommenders or setting a particular number of agreeing witnesses, as it is outlined in the following:

- Using social relationships for recommender selection. In social relationship-based reputation systems, such as in REGRET [34] where social relationships exist between peers, the cooperation relationship between a recommender and the trustee or between recommenders could be taken into account, so as to prevent some cases of collusion attacks. More specifically, in [34] the selection of recommenders is depended on their relationship with the trustee. Peers with a cooperation relationship with the trustee are not chosen for giving recommendations, to prevent a positive discrimination towards the trustee. Ashri et al. [63] suggest that a more elaborated relationship analysis should be incorporated in reputation models. They propose mechanisms for the identification and characterization of various relationships which may exist and change dynamically in an electronic marketplace (e.g. competition, collaboration), and consider the incorporation of these mechanisms in a trust model for e-commerce, in a way that for each identified kind of relationship between two entities specific rules are applied to their reputation estimation.

- Using a hash function of peers’ identifiers to determine which recommenders will send their recommendations to a specific peer. In Havelaar [57] recommenders always send their ratings to the same set of peers which are determined by a hash function of the sender’s identifier. In this way recommenders cannot determine to whom they will send their recommendations, thus, they cannot form coalitions. Furthermore, peers which frequently send unfair ratings for other peers can be easily detected by the receivers of the ratings.

- Using a number of pre-trusted peers (as in EigenTrust [30]) which collect recommendations and isolate unfair recommenders.
– Using a number of agreeing witnesses for recommendation verification. For example, Carbarun and Sion [64] use a mechanism referred to as threshold witnessing, where a set of witnesses provide service interaction feedback and estimate and sign reputation values for the interacting parties. For each new recommendation at least one non-faulty external observer is required to agree with it. The aim of this mechanism is to constrain the generation and modification of recommendations and reputation values and, thus, restrict ballot stuffing and bad mouthing attacks.

4.1.2.3. Tying a recommendation with a transaction (in case of transaction-based recommendations). So as to prevent recommendations related to fake transactions, and thus, ballot stuffing attacks. A simple, although loose, way to achieve this is by incorporating a timestamp in the recommendation, as in [37], where the tuple (destination, source, timestamp) is suggested to act as the key of the recommendation that will be used for storing and retrieving purposes. This timestamp (either referring to a transaction or to recommendation provision) can be used for verifying the originality of a transaction, although the authors do not describe such a mechanism. Kinadeter and Rothermel [53] propose the use of electronic payment schemes that will be combined with the reputation system and will allow the creation of an originality statement during the payment process which can be included in a recommendation in order to prove that a transaction regarding the recommendation really took place.

Srivatsa et al. [65] propose to bind a recommendation to a transaction through transaction proofs, built using a public key cryptography based scheme, so as a recommendation will be stored and used only if the source and the target of the recommendation really transacted with each other. Similarly with Srivatsa et al. [65], Singh and Liu [41] propose that two transacting peers exchange encrypted messages as proofs of interaction. Such encrypted messages can be used as a secure proof of a real interaction and are incorporated in the recommendation value.

4.1.2.4. Considering uncertainty and lack of evidence. Defense mechanisms in this category include:

– Estimation of confidence measures for each recommendation in case of opinion-based recommendations, and using them to weight the recommendation in the reputation metric. As pointed out in [66], confidence is an important concept because it can differentiate opinions established through a long-term experience and that through only a few transactions. The reputation model proposed by Theodorakopoulos and Baras [67] considers opinions as tuples consisting of two numbers: the trust value and the confidence value. The authors propose an algebra for estimating reputation values by combining these tuples in various ways. However, they do not consider how confidence values can be estimated. On the other hand, Sabater and Sierra [34] estimate a confidence value for each opinion-based recommendation based on the amount of available information the recommender has and on the deviation from the average of the ratings and use it to weight the recommendation. The more ratings are used for a recommendation and the more consistent these ratings are, the higher the confidence of the recommendation. In SORT [68] opinion-based recommendations are used based both on own experiences and recommendations from others. Each recommendation is weighed in the reputation formula using the size of transaction history between the recommender and the trustee and the number of recommenders which have contributed with their recommendations for its calculation. If the recommender had many transactions with the trustee, then its recommendation is more credible. Furthermore, when the opinion-based recommendation uses recommendations from a large set of peers, it is also considered more credible. In [69] the number of times a peer is asked to produce recommendations is proposed to be used as a confidence measure for the recommendations obtained by this peer.

– Incorporating uncertainty factors in the reputation estimation. Such an approach is proposed by Mogens et al. [70] where a reputation metric uses the number of transactions with unclear outcome as a measure against inaccurate recommendations attacks. In the object reputation system proposed by Ingram [45], where recommendations regarding an entity in a specific context are comments (notes), each peer maintains information about: (a) the relevancy that the recommendations received from a peer have with the specific context, (b) the accuracy of the recommendations regarding their factual correctness and (c) the peer’s opinion about how much it agrees with the non-factual claims in the recommendations. In this way the most useful (relevant, accurate and similar in preference) recommendations will be used. Uncertainty, fuzziness, and incomplete information in peer recommendations are also handled by Song et al. [28] using fuzzy logic inference rules. They propose FuzzyTrust, a reputation system for e-commerce, where the entities participating in an e-commerce transaction rate each other using fuzzy logic rules on local information concerning the transaction. In a number of reputation systems which model opinions such as probability distribution functions (e.g. [66, 71]), uncertainty in an opinion is modeled as the variation of the probability distribution function. Uncertainty is also explicitly modeled in [72] where each opinion includes an uncertainty factor corresponding to the lack of evidence to support a positive or negative opinion. Gutscher [73] deals with reputation systems uncertainty in more detail, considering conflicting opinions as an additional uncertainty factor which should be incorporated in the reputation metric.

– Allowing the trustee to provide the trustee with third party recommendations (certified reputation) regarding its transactions and incorporating them in the reputation metric. Such work is found in [74], where a peer A can ask its partner in a transaction, say peer B, to provide its recommendation to A regarding their transaction. Such recommendations can be sent by the peer A to future possible partners to help them estimate a reputation metric for peer A, especially when they can find no or little information about the trustee A from direct experiences or third party recommendations.
4.1.3. Incentive-based defense mechanisms

This category of defense mechanisms has to do with rewarding or punishing the recommendation behavior of peers; in this way, peers which provide fair recommendations are rewarded and those who do not provide fair recommendations or provide random opinions are punished. Related mechanisms found in the literature include:

4.1.3.1. Rewards (punishment) for providing honest (dishonest) recommendations. Papaioannou and Stamoulis [75] propose a mechanism which compares the ratings submitted by both transacting peers regarding the outcome of their mutual transaction. If the two ratings are in agreement they are taken into account in the calculation of each peer’s reputation, otherwise, both peers are punished. A peer under punishment is not allowed to transact with others. The same authors in [36] propose a side-payment scheme (monetary rewards for submitted feedback), which is used to give incentives to entities to report honest ratings. The reward is based on the correlation between the reports of different entities. They additionally propose the use of fixed monetary penalties in case two peers disagree about the outcome of a transaction between them.

The reputation model proposed by Fernandes et al. [51] and further enhanced by Kotsovinos et al. [76], incorporates both rewards and punishment depending on the participant’s behavior. More specifically, a credit balance is set up for every participant, which is credited with a reward for each opinion and debited for each recommendation query made by that participant. Honest peers are rewarded with credit discounts when asking for recommendations, but when judged as dishonest, they are punished with a probationary period during which no rewards are given to them for providing recommendations. The goal is that the punishment period is long enough to discourage participants from being dishonest several times, but not too harsh to disappoint first-time cheaters.

Another form of punishment is establishing higher transaction costs for sellers (in the case of P2P e-commerce systems), which collude with other peers to make a ballot stuffing attack with fake transactions. For example, in [77] a buyer will pay an incremental cost (reputation premium) for dealing with a seller of higher reputation and also a seller will pay to a central authority a fixed transaction cost for each transaction. The premium would tempt sellers to fake transactions to enhance their reputation and have thus higher revenues from their trades. So, the authors propose a premium estimation method that discourages sellers to fake transactions to increase its reputation, thus ballot stuffing attacks is prevented.

4.1.3.2. Payment scheme for buying and selling reputation information. As an incentive for providing honest recommendations, Jurca and Faltings [78] introduce a reputation mechanism which uses a payment scheme to motivate agents to report truthfully about their interactions’ results. They define a set of broker agents (called R-agents) whose tasks are buying and aggregating reports from other agents and selling back reputation information to them when they need it. In order to incentivize agents to share their reports truthfully, the proposed payment scheme guarantees that agents who report incorrectly will gradually lose money (during the process of selling reports and buying reputation information), while honest agents will not. Therefore, this mechanism makes it rational for an agent to report its observations honestly.

4.2. Defenses against inconsistent behavior

Defenses against inconsistent behavior comprise mechanisms that aim at restricting the impact of such behavior on the reputation estimation. Such mechanisms are the following:

4.2.1. Incorporating time in reputation estimation

Giving recent transactions more weight can be used to prevent traitors. This technique is used in [29], where two different weights are used in the aggregation of recommendations, one for older recommendations and one for new ones. A time decay function is proposed in [34,79–81] to weight recommendations according to the time they refer to, so that the more recent a transaction the higher the weight of a relevant recommendation. In the latter of the aforementioned works, time is also used to calculate a reliability measure for each reputation value. A similar idea, presented in [66], refers to the use of a forgetting factor (a number belonging in [0, 1]) which weighs recommendations according to their recency. The authors compare the use of a fixed forgetting factor with an adaptive forgetting factor which will depend on the current reputation value of the trustee. They conclude that an adaptive factor is more intuitive, as it mimics human reputation formation: bad behavior is remembered for a longer time than good behavior; also it takes a long time and consistent honest behavior to build a good reputation.

A time window can also be used for keeping the most recent recommendations and using them when estimating a reputation value. For example, in SORT [68] information about a service interaction is deleted after a certain period, and thus, only a limited number of recent transactions is kept and used for reputation estimation. Another example of using a reputation time window is found in [82], where the length of the time window determines the number of the most recent transactions with a peer that will be used for the calculation of the reputation value of the latter.

4.2.2. Incorporating context in reputation estimation

Various types of context can be used to assign weights to transaction ratings, depending on the specific transaction characteristics. For example, in a P2P e-business community, the size of a transaction could be used to weigh the feedback for that transaction as in [27,80]. This method can act as a defense in cases where a seller develops a good reputation by being honest for small transactions and then tries to make a profit by being dishonest for large and valuable transactions. Ramchurn et al. [83] tackle the same issue by incorporating the context of interactions and using it to filter out irrelevant interactions. In their model, peers negotiate on contracts, execute them and evaluate the outcome of the contract execution, i.e. whether it has been executed as agreed or not. Each peer stores a history of the agreed contracts, the context at the time a contract was negotiated and the outcome of the execution. The contract context may involve various types of information (e.g. seller location which implies that specific
4.3. Identity management-based defenses

Defenses in this category offer strong identification of peers, authentication of recommendation and reputation information as well as restriction of the number of identities a peer can have in a system. Their purpose is to enhance accountability and thus discourage or prevent the identity management-based attacks which have been presented in Section 3.3. Mechanisms in this category include:

4.3.1. Using unique digital identities

Unique digital identities aim at preventing impersonation. A Public Key Infrastructure [52] can be used for generating digital identities where each peer will be represented and uniquely identified by a pair of a public and a private key (e.g. [33]).

4.3.2. Using digital signatures

Digital signatures verify the authenticity of messages in order to prevent man-in-the-middle attacks. An example of such mechanism is the use of cryptographic mechanisms in order to preserve the integrity of reputation values or prevent peers from knowing the other peers’ recommendations, so as to avoid collusion and man-in-the-middle attacks [85]. Another example is the threshold witnessing mechanism proposed by Carbunar and Sion [64] which has already been discussed in Section 4.1.2; this mechanism incorporates the use of signatures in order to secure reputation values. Gutscher [86] presents a reputation model for open, decentralized systems which not only uses digitally signed recommendations but also integrates public key authenticity mechanisms. This integration enables the validation of public keys that sign recommendations and at the same time the validation of the reputation of an entity that claims that a public key belongs to a particular entity.

4.3.3. Restricting generation of new identities

Sybil attacks, whitewashers and ballot stuffing are discouraged by restricting new identity generation. Popular methods use:

- One persistent identity per ‘social arena’ (i.e. per online community such as forums or gaming network sites with varying degrees of anonymity need), so as to prevent whitewashers and Sybil attacks [87].

- Rate limits, where the rate at which a peer can generate new identities is limited by asking for example for each new identity the solution of a problem that requires computational power, bandwidth or storage space, as proposed in [15].

- Entry fees, where newcomers are charged a fee to join the system, so as to discourage Sybil attacks and whitewashers. Several techniques to do that are described in [87].

- A ‘few in a life time identities’, by enabling a few trusted companies to act as Certification Agencies (CAs), using blind signatures to preserve anonymity [61]. Each CA will only issue one identity per peer (the restriction on a CA issuing a second identity is time-based, e.g. one per year), so attackers will find it increasingly expensive to obtain more identities by other CAs. They will also find too slow the process of obtaining a second identity by the same CA.
Creation of new identities could also be prevented with the use of disincentives, by assigning for example, a low reputation to newcomers, and not allowing the reputation of a misbehaving peer to fall below this value [88]. This will discourage peers to discard their identity and enter the system with a new one. Another disincentive for Sybil attacks is found in the work of Seigneur et al. [89]. The authors suggest that for every recommendation an amount of the reputation of the recommender is transferred to the recommended entity. This strategy discourages Sybil attacks as peers with multiple identities cannot artificially increase the reputation of one of their identities, as the total reputation of a peer is shared between the identities it possesses. One can argue that it is unfair for a recommender to loose an amount of its trustworthiness when recommending another entity. For this reason, this strategy is limited to scenarios with many interactions between the recommenders; in such scenarios, the overall trustworthiness of each entity is large enough so there is no major effect on their trustworthiness when they transfer some of it to the recommended entity.

4.3.4. Controlled anonymity

This mechanism refers to concealing the identities of the seller and the buyer from each other, while making them known only to a marketplace administrator. The identities of malicious peers will not be able to attack and identify specific peers. The work by Dellarocas [31] uses controlled anonymity in order to avoid unfair ratings and negative discrimination in e-commerce.

4.3.5. Exploiting graph characteristics and properties of graph represented P2P networks

Such methods constrain Sybil attacks, as proposed in [90–93]. Yu et al. [90] propose the SybilGuard protocol which works with social networks represented as graphs where nodes are connected via strong human trust relationships. The SybilGuard protocol exploits properties of such graphs in order to bound the number of identities a malicious peer can create, based on the argument that malicious users can create many identities but few trust relationships. The same authors [91] propose the SybilLimit protocol for limiting further the number of Sybil nodes in fast-mixing social networks, based on the characteristics of graph represented social networks. Danezis and Mittal [92] propose SybilInfer which uses Bayesian Inference to label notes as honest and dishonest given a set of stated relationships between nodes. Finally, Cheng and Friedman [93] formulate a reputation system as a static graph, give conditions for sybil-proofness and propose reputation functions under these conditions.

5. Taxonomy analysis and roadmap for reputation system designers

In this section we present and discuss three tables (Tables 2–4) which summarize the presented defense mechanisms and map them to the attacks taxonomy depicted in Fig. 2:

- Table 2 summarizes the defense mechanisms against the unfair recommendations related attacks;
- Table 3 presents the defenses against the inconsistent behavior related attacks, and
- Table 4 presents the defenses against the identity scheme related attacks.

As one can see in these tables, the defense mechanisms have been classified according to the category of credibility factors that they concern. In these tables, we can observe that some types of attacks have no corresponding defense mechanisms under specific categories of credibility factors. This happens because each attack targets specific characteristics of a reputation system which, as described in Section 2, are related to specific classes of credibility factors. Therefore, the defense mechanisms against a specific attack target the affected classes of credibility factors rather than all of them. For example, the Man-in-the-middle attack is related to the lack of securing the integrity of recommendations and the weak authentication of recommenders and other mediators. Consequently, the defense mechanisms against this attack deal with preserving the integrity of recommendations (which belongs to the recommendation creation/content quality factors category of Section 2) and with authenticating and assessing the credibility of recommenders or other mediators (which belongs to the recommendation selection factors category). Therefore, the defense mechanisms developed against the Man-in-the-middle attack are all classified under these two classes of credibility factors as shown in Table 4 while the 3rd class of credibility factors (i.e., Reputation Reasoning) is not applicable (N/A).

Furthermore, it is worth to note that the study of Tables 2–4 reveals some ‘gaps’, where, to the best of our knowledge, there are no defense mechanisms addressing specific credibility factors. Such a case is the No Recommendations attack (Table 3), for which there are no defense mechanisms under Reputation reasoning category. This could mean that there is a need for further research. For example, the reputation estimation method could incorporate a mechanism for detecting and reducing the reputation of peers which do not give recommendations. So, new defense mechanisms against these attacks could be developed belonging to this specific class. Cells in Tables 2–4 revealing such gaps have been shaded in gray color and are worth to be carefully examined, so as to give useful directions for future work in defending against reputation system attacks.
<table>
<thead>
<tr>
<th>Attacks</th>
<th>Recommendation creation/content based DMs</th>
<th>Recommendation selection based DMs</th>
<th>Reputation reasoning based DMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unfair recommendations</td>
<td>• Incorporating uncertainty and lack of information in opinion-based recommendations [68, 34]</td>
<td>• Estimating recommendation reputation [27, 29, 55]</td>
<td>• Estimating the credibility of a reputation value [34]</td>
</tr>
<tr>
<td></td>
<td>• Incentives for honest recommendations [36, 51]</td>
<td>• Similarity based methods to identify unreliable recommenders from trustor’s point of view (filtering techniques) [57, 31]</td>
<td>• Incorporating a penalty factor and negative experiences sensitivity [84]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Estimating confidence for recommendations [56, 68]</td>
<td>• Aggregating a large number of recommendations [57]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Controlled anonymity [31]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Threshold witnessing [64]</td>
<td></td>
</tr>
<tr>
<td>Collusive badmouthing</td>
<td>Preventing peers to know the other peers’ recommendations via cryptographic mechanisms [85]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collusive reducing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reputation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collusive deceit, Ballot</td>
<td>• Tying recommendations with corresponding transactions [65, 41]</td>
<td>• Difficulty to change identities [87, 15]</td>
<td></td>
</tr>
<tr>
<td>stuffing</td>
<td>• Transaction costs [77]</td>
<td>• Threshold witnessing [64]</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3 – Defense mechanisms against inconsistent behavior based attacks.

<table>
<thead>
<tr>
<th>Attacks</th>
<th>Defense mechanisms (DM) classified according to the categories of credibility factors</th>
<th>Reputation reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recommendation creation/content</td>
<td>Recommendation selection</td>
</tr>
<tr>
<td>Inconsistent behavior</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Traitors</td>
<td>N/A</td>
<td>• Giving recent transactions more weight [29]</td>
</tr>
<tr>
<td>Service discrimination</td>
<td>N/A</td>
<td>• Incorporating context of recommendations in reputation estimation [37,80]</td>
</tr>
<tr>
<td>No recommendations</td>
<td>Rewards for honest recommendations [36]</td>
<td>• Incorporating a penalty factor and negative experiences sensitivity [84,80]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Detecting reputation variations [79]</td>
</tr>
</tbody>
</table>

### Table 4 – Defense mechanisms against identity scheme related attacks.

<table>
<thead>
<tr>
<th>Attacks</th>
<th>Defense mechanisms (DM) classified according to the categories of credibility factors</th>
<th>Reputation reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recommendation creation/content</td>
<td>Recommendation selection</td>
</tr>
<tr>
<td>Sybil attack, whitewashing/pseudospoofing</td>
<td>Self-recommendations with reputation transfer [89]</td>
<td>• Difficulty to change identities [87,15]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Exploiting graph characteristics of P2P systems [90-93]</td>
</tr>
<tr>
<td>Man-in-the-middle attack</td>
<td>Cryptographic mechanisms for securing recommendations [85]</td>
<td>• Cryptographic mechanisms for authenticating recommenders and mediators [86]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Estimating mediators credibility [44]</td>
</tr>
<tr>
<td>Impersonation</td>
<td>Unique digital identities [33]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unique digital identities [33]</td>
<td></td>
</tr>
</tbody>
</table>

### 5.2. Conflicts between desirable credibility features and required tradeoffs

The analysis of defense mechanisms brings to light several conflicts between desirable characteristics of P2P reputation systems which necessitate careful considerations in the reputation system designing process. Such desirable characteristics concern the credibility of the reputation system, scalability issues, ease of development, privacy issues, etc. The reputation system designer needs also to take into consideration the goals of the supported application and the characteristics of the application environment in order to choose the right tradeoffs when dealing with these conflicts. In the following we describe the aforementioned conflicts and the tradeoffs that have been found in literature.

#### 5.2.1. ‘Privacy’ vs. ‘Trust’

The inherent conflict between privacy and trust is described by Seigneur and Jensen in [94]. Privacy of a peer has to do with its ability to be anonymous and not to let other peers monitor its transactions and recommendations. However, for a reputation system to work, information about a peer’s identity, its transactions and provided recommendations needs to be monitored. The more information is linked to a peer’s identity (either a real-world identity or a pseudonym) the less its anonymity level and its privacy, and the higher the level of accountability and the reputation estimation credibility that can be achieved. A specific example of this conflict is found in [74] (which has been already been described in 4.1.2); although this mechanism results in higher availability of reputation information and in a quicker and more accurate reputation estimation, peers which give recommendations to the other entities surrender
their privacy with respect to how they value their partner's performance.

Seigneur and Jensen [94] propose tradeoff mechanisms based on the use of pseudonyms, i.e. virtual identities of entities, which should be linked to the transactions of an entity, considering that the ability to link a pseudonym with different transactions is required for building reputation. They consider linking of transactions with a specific pseudonym with the use of a digital signature-based entity recognition scheme for transactions [95]. They also propose a mechanism which links pieces of evidence to a specific pseudonym, and a protocol for disclosing to the trustor the minimum amount of evidence needed for a trust decision.

Deghaili et al. [96] propose a model for trust-privacy tradeoff in distributed systems which are organized as groups of agents. Their model aims to establish reputation-based trust relationships between agents before any interaction, with the least privacy loss possible. Each transaction between two peers requires disclosing pieces of information, for which a specific level of privacy loss is involved. Communication and disclosing of information between peers is achieved through the group administrators, and thus peers can choose to keep their anonymity in their transactions with peers from other groups.

The conflict between anonymity and reputation, which is a specific case of the conflict between privacy and trust, is analyzed in [97]. In this work two distinct 'identity infrastructures for P2P reputation systems' are compared under various threat conditions as follows: the first one uses a central trusted server which ties the real-world identities of peers to pseudo-identities; the second one is completely decentralized with each peer creating and managing its own identity, providing thus full anonymity. The first infrastructure is resilient to identity management-related attacks (e.g. Sybil and whitewashing attacks); peers, however, may find their privacy violated, as they are required to disclose information at a level that they may not find acceptable. The second reputation system protects peers' privacy and anonymity but it is quite vulnerable to Sybil and whitewashing attacks. The authors argue that the various techniques that can be used to defend against various attacks depending on the identity scheme used.

Privacy in reputation systems can be enhanced with the use of cryptographic mechanisms in order to hide the real identity of a peer and encrypt its communication with others [41]. Furthermore, the use of fungible reputation (which can be used as coins or tokens by an entity which hold various identities) is proposed in [98] as a way to enhance anonymity while preserving reputation.

5.2.2. 'Negative feedback sensitivity' vs. 'Robustness against collusive badmouthing'

When the reputation metric has a high sensitivity for negative recommendations, i.e. the reputation of a peer is significantly and quickly decreased when negative recommendations are encountered, then deceiving transactional behavior is discouraged. However, the reputation system becomes indefensible to bad mouthing and even more to collusive misbehavior. In this case, one or a combination of suitable defense mechanisms against unfair ratings could be used as a tradeoff mechanism, such as estimating the recommender's credibility, using filtering techniques to filter out dishonest recommenders or using incentive mechanisms for honest recommendations.

5.2.3. ‘Encouraging newcomers’ vs. ‘Preventing whitewashing’ via determining a default reputation value to newcomers

The initial trustworthiness value assigned to a newcomer should be high enough to encourage new peers to enter a P2P system; at the same time the initial value should not be so high in order to discourage malicious peers to misbehave and then enter the system with a new identity in order to escape bad reputation. A tradeoff for this conflict is presented in [88], where the minimum value of a peer's reputation cannot fall below the default value assigned to a newcomer, therefore peers do not have incentives to execute whitewashing attacks. Friedman and Resnick [87] study this issue comprehensively and state that distrust to newcomers is an inherent cost of the easy identity; thus, they discuss the use of entry fees for newcomers and the use of a single identity ‘per arena’ which could discourage whitewashers.

5.2.4. ‘Resiliency to oscillatory behavior’ vs. ‘Helping reputation restoration of previously misbehaving peers’ via monitoring changes in behavior

Taking into consideration previous transaction history and especially changes in behavior and incorporating these changes in the reputation metric mitigates oscillatory behavior but it has as a result that a misbehaving peer cannot easily restore its reputation. This may be unfair for honest peers which may unintentionally misbehave sometimes, e.g. by mistake, and thus decrease system utility. Mechanisms for monitoring recent behavior and rewarding it if it is honest and keeps being honest for a certain period of time (as in [51,76]) could help reputation restoration of previously misbehaving peers.

5.2.5. ‘Performance’ vs. ‘Accuracy’ via history size

The history size of transaction information and transaction evaluation information which is used for reputation estimation is very much related with the performance and the accuracy of the reputation system. The largest the history size the better the accuracy of the reputation metric but the smaller the performance, as both storage space and computation time needs are increased. Srivatsa et al. [65] propose the use of fading memories about recommendation information regarding a peer, aiming at keeping the right history sizes for achieving satisfying levels of both performance and accuracy. They propose recommendation data aggregation over intervals of exponentially increasing length, so as to keep a system defined size of values while obtaining these values from a larger number of reputation information.

5.2.6. ‘Performance’ vs. ‘Resilience to man-in-the-middle attacks’ via reputation information redundancy

Reputation information redundancy is employed by a number of reputation systems which estimate global reputation values, such as in [41,30] to facilitate resiliency to manipulation by malicious nodes. However, reputation information redundancy influences the performance of a reputation system,
as it poses increased requirements for storage and communication. As a tradeoff, Zhou and Hwang [40] propose the use of locality-preserving hashing to locate a small number of the most reputable nodes which are responsible for a peer’s global reputation. They argue that this mechanism provides low implementation overheads and good performance results.

5.2.7. ‘Considering only positive experiences, and thus countering badmouthing attacks’ vs. ‘Resilience to collusive deceit attacks’

When only positive experiences are considered in recommendations, then badmouthing does not affect the reputation of honest peers and thus the credibility of the reputation system. However, malicious peers can praise each other and increase each other’s reputation (collusive deceit attack) without giving the other peers the ability to complain about their dishonest behavior. When recommendations express degrees of satisfaction and not only positive experiences, both badmouthing and collusive deceit attacks can be more effectively prevented.

5.2.8. ‘Considering only negative experiences, and thus countering collusive deceit’ vs. ‘Resilience to badmouthing attacks’

Taking into account only negative experiences in reputation estimation counteracts the effects of collusive dishonest praises, but favors badmouthing attacks. As discussed in the analysis of the previous conflict, expressing degrees of satisfaction and not only negative experiences in recommendations, helps to more effectively prevent both types of attacks.

5.2.9. ‘Resilience to unfair recommendations via similarity measures’ vs. ‘Considering honest recommendations which do not comply with the majority of recommendations’

Recommendations which deviate from the average value of recommendations tend to be considered dishonest when similarity measures are used. They may be honest though, reflecting a sudden change in the trustee’s behavior or a discriminating behavior of the trustee towards the recommender (e.g. the trustee may choose to misbehave towards the recommender whereas it behaves honestly towards the majority of the other peers). In both cases the trustee behaves in a different way than usually, so a recommendation describing the specific behavior will deviate from the average recommendations, and it will not be taken into account, although it is honest. In order to deal with this case, similarity measures should take into account the similarity of a peer’s recommendations with other peers’ recommendations regarding not only the specific trustee but also the other recommended peers. When a peer’s recommendations deviate from other peers’ recommendations regarding the majority of recommended entities and not only a specific entity, then it could be considered safe to exclude this peer’s recommendations when estimating reputation values for other peers.

However, there is also the case of discriminatory recommendation behavior where a peer may recommend most entities similarly with the average of other peers’ recommendations but give unfair recommendations only to one other entity. Therefore, appropriate tradeoffs still need to be considered, such as those in [31,56], which have been presented in 4.1.1.

5.2.10. ‘Encouraging recommendation provision via rewards for recommendations’ vs. ‘Preventing random recommendations’

Some reputation systems (e.g. [51]) reward peers which give their recommendations for their partners; this is an incentive especially for high reputable peers who may not want to give good recommendations for other peers and increase their reputation. However, when such rewards are given, peers may be tempted to give dishonest opinions. To achieve a tradeoff between incentives for recommendations and preventing random recommendations, Fernandes et al. [51] and Kotsovinos et al. [76] propose that punishment mechanisms for dishonest recommendations should be deployed to ensure that rewards are given only for honest recommendations.

5.2.11. ‘Incentives for honest recommendations via credit-based reward/punishment mechanisms’ vs. ‘Ease of development’

Credit-based mechanisms incorporated in reputation systems, such as [51,85], provide incentives for honest recommendations. However, they require increased communication and implementation costs as the credit balance of each peer should be kept and updated based on complex economic models in a decentralized manner. Work for achieving right tradeoffs in decentralized systems is in preliminary stages, as research in the field of credit-based incentive mechanisms for reputation systems has been mostly based on centralized architectures. Gupta and Somani [32] try to achieve such a tradeoff in P2P systems using a semi-centralized architecture. Their proposal for a reputation management framework uses special peers to calculate and store the credit balance of other peers aiming to a light-weight and secure credit-based reputation system with low computational and communication overheads.

Three more works proposing the implementation of credit mechanisms for P2P applications that are worth to be mentioned, although not specifically related with reputation systems, are the ones found in [99–101]. Vishnumurthy et al. [100] propose a framework for a distributed credit based mechanism for P2P applications, where a credit balance is kept for each peer depending on its levels of service consumption and service contribution in the system. Credit accounts of peers are managed by a set of specialized nodes. Belenky et al. [99] propose the implementation of e-cash for the fair exchange of services in P2P applications. The proposed implementation which claims to mitigate communication and computation costs is based on a centralized architecture. Garcia and Hoepman [101] present a completely decentralized currency implementation for P2P and grid systems, where credit accounts of peers are managed by a pre-determined, but random, selection of nodes.

5.3. Roadmap for reputation system designers

We believe that the presented taxonomy and analysis can be used as a reference guide for building an attack resistant P2P reputation system. Thus, reputation system designers can benefit from our work by following the steps below:
1. Identify the specific attacks that can threaten the reputation system at hand by considering the characteristics and requirements of the specific application (i.e. application area, application participants, desirable anonymity level, social relationships, etc.) and by taking into account the attacks taxonomy provided in this paper.

2. Choose suitable candidate defense mechanisms to implement. The analysis of available defense mechanisms against those attacks presented in Section 4 can help in this process.

3. Make final decisions regarding: (a) the defense mechanisms to implement and the associated implementation cost, and, (b) the characteristics of the reputation model. The developer can be assisted in this process by taking into account:
   - Table 1 which relates the reputation system characteristics with the categories of credibility factors;
   - Tables 2–4 which map defense mechanisms to the three categories of credibility factors of a reputation system;
   - The conflicts between the desirable characteristics of reputation systems and/or the defense mechanisms, as presented in Section 5.2, in order to achieve suitable trade offs.

In addition to the above steps, and based on the presented analysis of attacks and defenses, we summarize the most important issues that can further assist a reputation system designer and which need to be taken into account (when applicable) in order to build a P2P reputation system that is resilient to attacks:

- **Use of strict registration and authentication policies.** Such policies (e.g. public key infrastructure and sophisticated public key policies, as in [41]) enhance accountability of actions, and thus resistance to individual and collusive deceit. Encryption techniques can also help against privacy breaching attacks, weak authentication based attacks and reputation information integrity violations.

- **Recommendation credibility.** This can be enhanced by linking a recommendation to a related transaction in case of transaction-based recommendations or by estimating a confidence measure based on the amount and variability of information in case of opinion-based recommendations.

- **Recommender selection.** It should be free of bias, e.g. social relationships should be taken into account in social network-based reputation systems. Such selection can be based on recommendation relevance or on credibility, by either using similarity filtering techniques or by incorporating a recommendation reputation measure.

- **Evaluation of a peer’s trustworthiness.** This can be based on a dynamic trust metric, which incorporates a number of factors, such as:
  - time, giving more relevance to recent transactions;
  - the number of recommendations, as a measure of confidence that can be placed on the estimated reputation value;
  - sensitivity on negative ratings and on sudden changes of transactional behavior;
  - recommendation trustworthiness related to each recommendation.

- **Honest recommendation behavior.** This could be motivated with suitable incentives or by punishing dishonest recommenders.

### 6. Concluding remarks

Motivated by the fact that research in the area of decentralized reputation systems lacks an exploration of realistic adversary models which can include sophisticated and coordinated attacks, we provided a thorough analysis of the threats towards reputation systems through a taxonomy of reputation attacks. Specifically, we presented the main categories and subcategories of the attacks, followed by the defense mechanisms which are found in the literature. The defense mechanisms have been classified in a way that correlates them with the specific attacks. The mappings between attacks and defense mechanisms are presented in respective tables.

An additional contribution of our work is the identification of areas where there is a need for further research work in developing new defense mechanisms against reputation attacks. These areas refer to specific credibility factors, the right treatment of which can enhance the robustness of a reputation system against specific attacks. They are indicated in Tables 2–4 as empty cells shaded in gray.

The presented results have been utilized for the construction of a roadmap for reputation system designers on how to build attack resistant P2P reputation systems for robust P2P applications. The basic steps and issues comprising the roadmap are described in Section 5.3.

As there is a clear need for additional work on reputation systems for P2P applications which are resistant to various threats, we will continue working in this direction. Our future work plans include benchmarking of P2P reputation systems as well as engineering of credibility requirements in order to use them for building credible reputation systems for P2P e-community applications.

### References


