Survey on social reputation mechanisms: Someone told me I can trust you

Thomas Werthenbach
Delft University of Technology
Delft, The Netherlands
T.A.K.Werthenbach@student.tudelft.nl

Abstract

As internet availability remains to spread across the globe, online trust has become an increasingly relevant subject. One way to obtain a measure of trust is through reputation mechanisms, which record one's past performance and interactions to generate a reputational value. We observe that numerous existing reputation mechanisms share similarities with actual social phenomena; we call such mechanisms 'social reputation mechanisms'. The aim of this paper is to discuss several social phenomena and map these to existing social reputation mechanisms in a variety of scopes. First, we focus on reputation mechanisms in the individual scope, in which everyone is responsible for their own reputation. Subjective reputational values may be communicated to different entities as in the form of recommendations. Secondly, we discuss social reputation mechanisms in the acquaintances scope, where one's reputation can be tied to another through vouching or invite-only networks. Finally, we present existing social reputation mechanisms in the neighbourhood scope. In such systems, one's reputation can heavily be affected by the behaviour of others in their neighbourhood or social group.

Keywords: Social reputation mechanism, trust, trust graph

1 Introduction

As internet availability remains to spread across the globe, online trust has become an increasingly relevant subject. The COVID-19 pandemic has shown that, in time of crises, the online news and social media usage increases [1], increasing the risk and impact of misinformation. As such, it is commonly known that governments have attempted to control news media to spread propaganda in the past. Additionally, research shows that individuals getting their news from social media are often more likely to belief conspiracy theories [2]. Such matters raise the relevant and contemporary question: who to trust?

In an automated setting, the trust measure is often extracted from one's reputation. Their reputation may be calculated through the amount of 'good' work one has performed, or the reputation of their direct peers. We call systems performing such calculations *reputation mechanisms*. Many reputation mechanisms have been proposed and evaluated [3–10]. The core components of reputation mechanisms may vary greatly, e.g. it may assume that entities have a fixed initial identity or that some entity i sending some entity j a message provides a proof of personhood for entity j. However, the common purpose of reputation mechanisms is to provide some measure of benevolence or trustworthiness.

The overall scope of this paper is focused on *social reputation mechanisms*. Such reputation mechanisms are a virtual reflection of genuine social phenomena, such as vouching or familial relationships. We provide a survey in which we have rigorously reviewed social reputation mechanisms by exploring various social concepts and mapping these to existing reputation mechanisms.

Through the course of this paper, we gradually increase our scope and consider social reputation mechanisms based on social phenomena on an increasingly larger scale. First, we discuss the individual level. In this scope, no two persons necessarily know each other initially and everyone's reputation is based solely on the work they perform or the quality they provide. Secondly, we consider one's acquaintances. Existing social ties and vouching are concepts which may transpire in this space. Lastly, we discuss phenomena occurring in one's direct neighbourhood. For instance, the neighbourhood in which you live may affect your reputation to both members outside and inside that neighbourhood.

First, we provide more background on the importance and relevance of creating trust and reputation mechanisms in section 2. Section 3 provides formal definitions and data structures, which we use to generalise the mathematical foundations of reviewed mechanisms in order to reduce the usage of varying mathematical models across the different reputation mechanisms. Section 4 considers entities individually, and rigorously discusses different reputation mechanisms based on social phenomena within this scope. Section 5 continues exploring social concepts in the acquaintances scope and their associated social reputation mechanisms. The last scope, neighbourhoods, is discussed in section 6. A brief overview of all discussed social reputation mechanisms can be found in table 1.

2 Background

Shaping trust in the online world, arguably the telos of all reputation mechanisms, is a hard challenge, which has been studied as early as 2002 [30–32]. As the space of defense mechanisms gradually evolves, so does the space of attack

Table 1. Overview of all social reputation mechanisms reviewed in this paper, as well as associated work.

Year	Mechanism	Reputation model	Related
1999	**PageRank** [4]. *Type: individual.* Assumes important websites are likely linked to from other websites. Links are quantified iteratively until stationary state.	$R_i = c \sum_{v \in B_i} \frac{R_i}{\lvert N_i \rvert}$	[11–14]
2006	**GroupRep** [9]. *Type: neighbourhood.* Users form natural groups. The group reputation is assumed when there has not been sufficient direct interaction.	Given utility u_{ij} and cost c_{ij} from i to j: $R_j = \frac{c_{ij} - u_{ij}}{c_{ij} + u_{ij}}$	[15–19]
2009	**IPGroupRep** [10]. *Type: neighbourhood.* Adopts IP-based groups and aggregates spam detection feedback for reputation values.	Given positive feedback r and negative feedback s: $R_i = \frac{r+1}{r+s+2}$	[20–22]
2010	**ARRep** [3]. *Type: individual.* Leverages direct experiences with recommendations.	$R_{ij} = \alpha \cdot R^D_{ij} + (1-\alpha) \cdot R^R_{ij}$	[23–27]
2011	**Trust by Association** [7]. *Type: acquaintances.* Invite-only network; reputation of the invitee directly affects inviter.	Given some underlying reputation mechanism U: $R_i = (1-\alpha)U_i + \alpha \frac{\sum_{j \in N_i} U_j}{\lvert N_i \rvert}$	[28]
2012	**Souche** [6]. *Type: acquaintances.* Frictionless vouching and assumes all benevolent users are member of a giant connected component.	Given a giant connected component (GCC), which growth is limited for each time interval: $R_i = \begin{cases} \text{Trusted}, & \text{if } i \in GCC \\ \text{Not trusted}, & \text{otherwise} \end{cases}$	
2015	**SocialTrust** [8]. *Type: acquaintances.* Prefer friends over strangers. Relies on reputation of strangers if no friend available.	Entity's reputation is modified based on the rating of the other party and their impact factor T_i: $T_i = \beta \frac{R_i}{R_{max}} + (1-\beta) \frac{D_i}{D_{max}}$	[29]
2022	**MeritRank** [5]. *Type: individual.* Defines set of strategies to make Sybil prone reputation mechanisms Sybil tolerant.	Transitivity decay, connectivity decay and epoch decay applied on existing reputation mechanisms.	

possibilities. For instance, people are getting more aware of the risk of the internet and start to become sceptic towards (spam)mails, causing scammers to invent more intelligent and sophisticated scams [33]. Another example of the need for online trust is in the world of e-commerce, where criminals are actively attempting to swindle innocent users on large e-commerce platforms, like eBay [34].

Nowadays, this responsibility of creating trust cannot be entrusted to private corporations. In recent events, Alphabet Inc. has been fined €220 million by French authorities for abusing its dominance in the advertisement industry. The French government has accused Alphabet Inc. of promoting their own advertisements over their competitors' in their search engine, Google. Furthermore, in 2019, Google has been fined €1.28 billion by the European Union on similar charges [35]. Google's dominance in the advertisement industry and the abuse of this position manifests their absolute control over the ranking of advertisements and online resources, incentivizing one to dispute their role in creating online trust. This case shows a typical example of the Red Queen hypothesis, which, in the e-commerce setting, states that companies must constantly adapt/evolve to stay ahead of their evolving competition [36]. Such online wars only help in creating distrust between different parties, strengthening the need for widely accepted trust mechanisms.

Exploiting social phenomena for the purpose of creating trust in online settings has previously been considered with the proposal of a novel peer-to-peer file-sharing system, named TRIBLER [37]. TRIBLER is a peer-to-peer file-sharing system, which introduces social ties to incentivize users not to misbehave at the expense of their friends, partners or community. TRIBLER suggests the usage of public and private keys as an authenticational method for recognizing previously encountered users in the anonymous peer-to-peer environment, enabling users to keep track of benevolent and malicious interactions.

3 Definitions

This section provides the formal definitions of various concepts and data structures we use in the description of existing social reputation mechanisms.

Entity – The notion of an entity encapsulates any type of instance which may participate in the network employing the underlying social reputation mechanisms. For example, an entity may be a real person, but could also be a computer.

Reputation mechanism – We adopt the definition of reputation mechanism as formulated by Swamynathan: "*A reputation mechanism collects, aggregates, and disseminates feedback about a user's behavior, or reputation, based on the user's past interactions with others*" [38]. In other words, a reputation mechanism processes feedback received from all entities participating in the network to cumulatively calculate a subjective or global reputation value for each entity.

Trust – In the context of computing systems, we may adopt the definition of trust as formalized by Saputra: "*Trust is a Trustor's level of confidence in regard to the ability of a Trustee to provide expected result in an interaction between Trustor and Trustee*" [39], where a trustor is the party which receives some service and the trustee is the party entrusted with performing or providing the trustor with a certain service or resource. In other words, trust is the certainty at which entity A (trustor) believes that entity B (trustee) is able to provide them with some service. More formally, trust is defined as a weighted directional relation $(i, j, v) \in E$ between two entities $i, j \in N$ and $v \in \mathbb{R}$, where N is the set of all entities, E is the set of all directed relations between two entities and v is the trustworthiness value assigned by some entity i to some entity j.

Trust graph – Trust relations as defined previously can be aggregated in a directed graph. We call such graph a *trust graph*, or alternatively *social graph*. This graph is defined by the tuple (V, E), where V is the set of entities and E is the set of trust relationships, also referred to as edges. Such a trust graph often facilitates the necessary structural foundation. More specifically, we say that if some entity i which has had sufficient (in)direct interaction with some arbitrary entity j, such that $j \in N_i$ and $\exists(i, j, v) \in E : v \in \mathbb{R}$, where N_i is the called a *trust set*, consisting of entities with whom entity i has had sufficient interaction with to assess their trustworthiness, depending on the underlying reputation mechanism. Furthermore, entities can occur in multiple *trust sets*, but no entity can contain itself in its trust set: $\forall i \in N : i \notin N_i$. Additionally, all entities occur at most exactly once in every trust set, such that $\forall i \in N : \{\forall j, k \in N_i : ID(j) = ID(k) \Leftrightarrow j = k\}$, where ID is a deterministic implementation-specific function capable of identifying individual entities. Note that the prior implies that $\forall(i, j, v) \in E : i \neq j$. We argue that every directional relation in the graph is unique, such that $\forall(i, j, v), (k, l, w) \in E : \{(i = k \wedge j = l) \Leftrightarrow (i, j, v) = (k, l, w)\}$. Finally, all entities occur exactly once in a *trust graph*: $\forall i, j \in N : \{ID(i) = ID(j) \Leftrightarrow i = j\}$. An example of a trust graph can be found in figure 1.

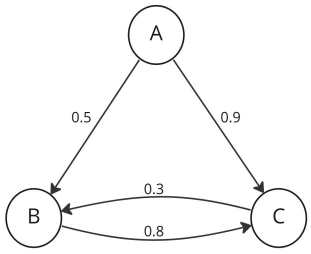

Figure 1. Example trust graph representing trust relations between entity's A, B and C. In B's perspective, C has a trust/reputation value of 0.8, implying that $(B, C, 0.8) \in E$. The reputation value is calculated and interpreted by the underlying reputation mechanism. Furthermore, B's trust set corresponds to $N_B = \{C\}$ and A's trust set to $N_A = \{B, C\}$.

Sybil attacks – The Sybil attack [31] is a well-known attack used against reputation mechanisms. Many reputation mechanisms are unable to distinguish original entities from their copies [40]; a weakness abused by the Sybil attack. An adversary may employ the Sybil attack to increase its own reputation through the instant creation of virtual entities, such that they may enjoy the benefits of high reputations. The method used to increase one's reputation using 'Sybil entities' depends heavily on the implementation details of the underlying social reputation mechanism. In 2011, Seuken et al. have shown that under specific circumstances, there exists a passive strongly beneficial Sybil attack [41]. In such an attack, a malicious entity can obtain an infinite gain with minimal effort.

4 Individuals

In the individual scope, no two entities have any initial subjective reputation value of each other and all reputations are based on the work the entities perform. However, once an entity has attained the trust of some other entity, it might propagate this trust value to peers, depending on the underlying mechanism. A physical social phenomenon resembling such situation is a networking event. During networking events, no two people have any initial measures of trust of each other, but any two people may grow to trust each other through reciprocity. The gradual creation of such trust relations may be used to form a trust graph. The assessed trustworthiness may then be shared throughout one's 'network', such that someone can obtain a reputation value of an entity, which whom they did not have direct interaction.

While direct experience with an entity is the most reliable metric to assess the trustworthiness of an entity [42], sociological research has found that reputational values are often spread through gossip [43]. Recipients of such reputational values have been shown to use these to selectively interact

with cooperative rather than selfish individuals. An example of a reputation mechanism adopting this social behaviour in an online setting is ARRep [3].

ARRep

ARRep (adaptive and robust reputation mechanism) [3] is a social reputation mechanism which leverages direct experience with reported experiences from other entities. While ARRep is proposed for usage in a peer-to-peer environment, the resemblance with the social phenomenon as depicted previously is vivid. Furthermore, ARRep applies heuristic for improving the accuracy of reported experiences, by giving more weight to entities who have had more experiences.

Given some entity i assessing the trustworthiness of some entity j, j's overall reputation value R_{ij} can be calculated according to:

$$R_{ij} = \alpha \cdot R_{ij}^D + (1 - \alpha) \cdot R_{ij}^R$$

where R_{ij}^D represents the reputation value extracted from i's direct experience with j, R_{ij}^R corresponds to the reputation value extracted from the recommendation of peers, and α represents the confidence factor of i's direct experience. For some threshold $M > 0$, α is equal to the ratio between the number of experiences and M while the number of experiences is lower than M, otherwise $\alpha = 1$. The value of R_{ij}^D corresponds to:

$$R_{ij}^D = \frac{\sum_{k=1}^{n_{ij}} (\lambda^{n_{ij}-k} \cdot ex_{ij}^k)}{\sum_{k=1}^{n_{ij}} \lambda^{n_{ij}-k}}$$

where n_{ij} is the total number of interactions between i and j, λ is some decay value such that $0 < \lambda \leq 1$ and ex is a function returning either 1 (good) or 0 (bad) depending on the experience of interactions between i and j from i's perspective. Moreover, the recommended reputation value R_{ij}^R is calculated, such that:

$$R_{ij}^R = \frac{\sum_{i \neq k} (C_{ik} \cdot R_{ik}^D \cdot \eta^{1/n_{kj}})}{\sum_{i \neq k} C_{ik}}$$

where η denotes some value $0 < \eta \leq 1$ and C_{ik} corresponds to the recommendation credibility based on the similarity between entity i and the recommender k (see [3] for details).

During evaluation, it was found that ARRep outperforms existing work [23] in a number of attacks for which peer-to-peer networks are susceptible. More specifically, ARRep has shown to performs better in *on-off attacks*, *bad mouthing attacks* and *collusive cheat attacks*.

There exist several reputation mechanisms similar to ARRep, focused on the same principles of combining direct experience with recommendations [23–27]. Continuing on the phenomenon in which reputation may be passed on through gossiping, an example of a reputation mechanism which directly applies this, is PageRank. PageRank uses the number of references an entity receives to determine its reputation compared to others. This behaviour is again very similar to that during networking events. PageRank has been used for assigning reputation values in social networks [44] or to measure academic reputation through citation graphs [45].

PageRank

In the early ages of the internet, Google was among the first to adopt a reputation mechanism. Larry Page, Google's co-founder, introduced PageRank [4]: an algorithm used to rank search engine results based on relevance. While PageRank might no longer be Google's only reputation mechanism, it is the basis of numerous other reputation mechanism [11–14].

PageRank considers the internet as a network of web pages connected through their links. If many pages link to another page, it has a higher reputation and therefore a higher 'rank' on the search results page. PageRank's algorithm employs the usage of rounds: initially, every page has the same amount of 'rank'. Every subsequent round, the rank flows uniformly distributed over all outgoing links to other web pages (see figure 2). Once the network reaches a stationary state, i.e. the rank does not change anymore, extracting the amount of rank per web page is trivial. One may note that this algorithm shows high similarity to finding the limiting probabilities of a Markov chain.

Let A be a matrix such that $\forall(i, j, v) \in E : A_{i,j} = \frac{1}{|N_i|}$. Note that the value v is not used by PageRank as it utilizes the notion of global reputation, i.e. the reputation is equivalent from all perspectives. Let R the reputation value of web page i, such that:

$$R_i = c \sum_{v \in B_i} \frac{R_i}{|N_i|}$$

where B_i is the set of states $\{j \in N \mid i \in N_j\}$ and c is a factor used for normalization, ensuring the total amount of 'rank' remains constant. When R reaches a stationary state, i.e. it does not change anymore, it is an eigenvector of matrix A, such that $A = cAR$. However, if the trust graph takes the shape of a directed cyclic graph, loops with no outgoing edges may occur, causing the accumulation of rank over time. To tackle this issue, Page a new formula for reputational values R' of web page i, such that $R_i' = R_i + cS_i$, where $||R'||_1 = 1$ and S_i is a vector of web page i which corresponds to the rank originating from each page. As we have that $||R'||_1 = 1$, c must be reduced when S is an all-positive vector, implying that c is a decay factor.

The original version of PageRank as described above is prone to Sybil attacks, as has been shown in many studies [46–49]. Such an attack would introduce many new entities who all link to the attacker, thereby increasing its reputation. This process is also known as 'link farming' [48]. The original PageRank algorithm does by itself not contain any defense mechanisms against Sybil attacks.

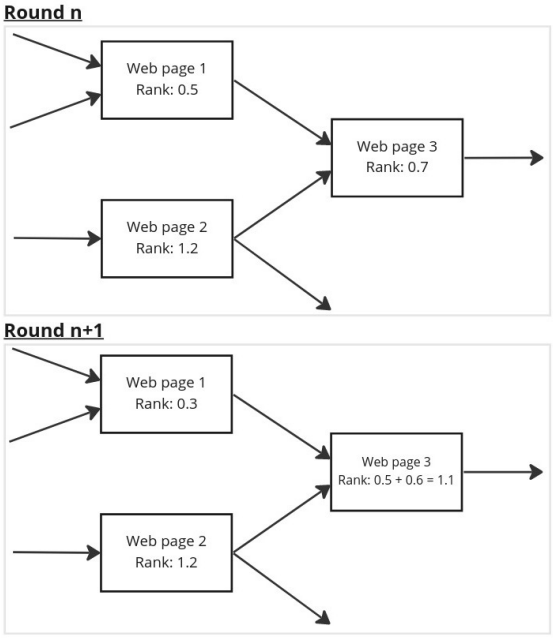

Figure 2. PageRank execution in which the edges represent references. In round n, pages (or entities) have a specific amount of rank. In round n+1, the rank is propagated over the outgoing edges. Web page 3 passes the 0.7 rank over its only outgoing link while receiving 0.5 from web page 1 and $\frac{1.2}{2}$ from web page 2. Web page 1's rank is lowered and web page 2's rank is stationary. This process continues until the amount of rank for all pages becomes stationary.

PageRank is part of the family of *symmetric reputation mechanisms*, which are generally prone to Sybil attacks [40]. In such mechanisms, the reputation of an entity does not depend on its identity, but only on the topology of the trust graph. An example of a defense against Sybil attacks in such scenarios, is MeritRank, which wraps existing social symmetric reputation mechanisms and adds additional constraints, providing these mechanisms with Sybil attack tolerance.

MeritRank

MeritRank [5] is a novel reputation mechanism which main goal is to bound the gain of Sybil attacks. That is, MeritRank does not attempt to solve Sybil attacks, but merely defines a number of strategies towards tolerating them. Furthermore, MeritRank generically assumes the existence of an underlying implementation for communication and reputation calculation using a 'flow-based' network, much alike the implementation used by PageRank.

Trust graphs satisfying MeritRank's constraints are shown to be Sybil tolerant. That is, for some value $0 < c < \infty$ and Sybil attack σ_S, the following holds:

$$\lim_{|S|\to\infty} \frac{\omega^+(\sigma_S)}{\omega^-(\sigma_S)} < c$$

where S is the set of Sybils, ω^+ is a function returning the gain for a Sybil attack and ω^+ is a function returning the amount of loss for a Sybil attack. By defining certain properties for trust graph, MeritRank is capable of bounding the amount of gain an attacker can get from attacking the network. Such an attack is also known as a weakly beneficial Sybil attack [50], which contrasts an attack where an adversary can obtain infinite gain, also known as a strongly beneficial Sybil attack. The constraints which MeritRank poses the trust graph are relative feedback/reputation, connectivity decay, transitivity decay and epoch decay.

The aforementioned constraints are a set of intuitive measures to bound the gain of an adversary. Relative feedback limits the amount of reputation an entity can give to some other entity by its own degree. More specifically, the updated function for assigning reputation is defined as:

$$\bar{w}(i,j) = \frac{w(i,j)}{\sum_{k\in N_i} w(i,k)}$$

where w is the original function for assigning reputation. Note the sum of reputation/feedback an entity assigns to its neighbours consistently equals 1. Transitivity decay defines a probability α which is equivalent to stop a random walk (see the Random Surfer model [4]) for reputation determination for any given entity. Furthermore, connectivity decay defines a constant $0 \leq \beta \leq 1$ and ratio t, such that if for some entity i (transitively) connected to some entity j through some entity k for at least the ratio t of all possible paths, $(1-\beta)$ serves as a punishment factor for decreasing the reputation of the entity j in i's perspective. The connectivity decay constraint's main purpose is to identify and punish separate components. Lastly, the epoch decay defines a constant γ, which indicates the reputation decay with each epoch of the graph, incentivizing entities to keep performing work to receive reputation.

MeritRank has been evaluated on all constraints separately. It has been shown that "transitivity decay and connectivity decay can provide a desirable level of Sybil tolerance" [5]. On the other hand, it was found that epoch decay, when naively implemented, may prefer new reputation assignments over existing reputation assignments. As aforementioned, MeritRank does not provide resistance against Sybil attacks, but accepts their existence and introduces a number of possible strategies towards bounding the maximum gain an attack may muster.

Individually-based social reputation mechanisms are often the most prone to Sybil attacks, as there exists no other external notion on which to base reputation calculations. While MeritRank proposes a number of strategies towards tolerating such attacks, it recognizes their existence and its inability towards preventing them. Arguably the most effective defense against Sybil attacks is the usage of fixed identity's and disabling the arbitrary creation of new virtual entities.

An example of such fixed identity is the European Digital Identity [51], which will enable EU residents to claim a single online identity. An external party verifying or providing an entity's identity is said to be the only way of preventing Sybil attacks [40], as identities cannot instantly be created without the external party's permission and verification.

5 Acquaintances

In the scope of acquaintances, we consider social reputation mechanisms which rely on the existence of real relationships between entities. By leveraging these existing relationships, one may strengthen the defenses of online social reputation mechanisms. An example of a social phenomenon leveraging existing relationships is *vouching*: *"to be able from your knowledge or experience to say that something is true"* [52]. In the context of reputation mechanism, vouching may generally be used as a method of putting one's reputation at stake. More specifically, in the case where some person (the *voucher*) has vouched for someone else (the *vouchee*), while this vouch was misplaced, the voucher loses their credibility. As a voucher willingly puts their reputation at stake for the vouchee, it makes one believe that the voucher has had prior external experience with the vouchee.

In recent years, the government of the United Kingdom has composed a rigorous guide as how to use vouches in daily-life situations [53]. It describes how people can use vouching for verifying one's identity. For instance, a parent has the ability to vouch for their child's identity. They know their child well and are certain of their child's identity, inducing no risk of vouching for them.

An example of a social mechanism employing vouching is Souche, which can be deployed on online social networks for protecting real users against fake accounts, often created for malicious purposes, such as spamming.

Souche

Souche [6] is a vouch-based reputation mechanism developed partially by Microsoft[1]. Its main goal is to quickly be able to distinguish between legitimate and illegitimate users in the context of online social communities, and to slow down any malicious undetected users. Souche has been evaluated in simulations utilizing large anonymized email and Twitter[2] datasets and has been shown to accurately identify 85% of legitimate users in an early stage. Furthermore, Souche can relief users of periodic humanity checks, such as CAPTCHAs, by only performing a CAPTCHA upon registration.

Souche's main means for creating relationships between entities, i.e. users, is through implicit vouching. Such process takes place through by considering regular activities as vouching. As such, Souche defines a vouch through emails by the conversation between two users, i.e. both users have written each other at least two emails for a conversation to be considered a vouch. Moreover, when modelling such approach to large datasets, it was found that a *Giant Connected Component* (GCC) starts to take shape. Such a GCC is a large trust graph which contains 93% of all users for the e-mail dataset, where the remaining connected components are orders of magnitude smaller than the GCC. Souche crowdsources the detection of malicious accounts, by assuming that malicious accounts are not included in the GCC.

Souche defines a quota q_i for each entity i to determine whether an entity is allowed to vouch for some new entity. Every unit of time, this quota grows with rate r. An entity is allowed to vouch for some other entity when their quota is larger than 1. Naively, the quota can be defined as:

$$q_i = (1 + r)^{t-b_i} - c_i - 1$$

where t is the current time, b_i is the time at which entity i joined the network and c_i indicates the number of entity i has already vouched for. However, in order to approach the growth rate with which online social networks grow, growth rate r should be configured to have a small value, such as 0.001 where the time interval equals 1 day. This implies that users are unable to vouch for any other users during their first registered year. To tackle this issue, Souche divides the GCC trust graph in subtrees, starting at the leaves, i.e. entities with no outbound vouches. An example of such a subtree can be found in figure 3. These subtrees have a size of approximately 50 entities and have a single root. Within subtrees, entities can freely use the cumulative quota. More specifically, entity i of subtree T_i can vouch for some other entity when $\sum_{k \in T_i} q_k > 1$. In order to account for the usage of shared quota, the definition of quota is finalized to:

$$q_i = (1 + r)^{t-b_i} - c_i - d_i - 1$$

where d_i represents the quota used by other entities to retain the total balance of quota within the network. Note that, due to the exponential growth of quota, older entities are assumed to be more trusted vouchers.

Other than sharing quota, the subtree data structure serves another purpose, namely that of assisting in the detection of malicious entities. While Souche itself does not focus on malicious entity detection, given an existing detection implementation, Souche can assist by marking an entities's parent, siblings or descendants as suspicious. Another defense against malicious entities is the limited quota per time interval, preventing adversaries from vouching for other adversaries. Smaller trees will result in less available shared quota for malicious entities to claim. Finally, Sybil attacks may also suffer from these features.

Another example of a study applying a vouching-based mechanism has been employed by the CloudSurfing platform [54]. This approach implements a more explicit method of vouching, requires more manual user interaction, and does not

[1] https://microsoft.com/
[2] https://twitter.com/

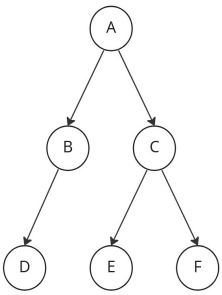

Figure 3. Example of a Souche subtree. In this particular example, A has vouched for both B and C and C has vouched for E and F. Note that if C were to be exposed as a malicious entity, it is evident that at least A, B, E and F should be further investigated, as they share a close relation to C.

protect users from malicious and potentially fake entities, but is used as a rating for hosts on the CloudSurfing platform.

On the other hand, there exist other social phenomena leveraging existing relationships which have been translated to social reputation mechanisms. One such example includes the usage of invitations. In an offline setting, invitations are often used to invite people to participate in a certain event. This social behaviour has been studied and integrated as a core component in social reputation mechanisms, such as *Trust by association*, which combines the usage of invitations with a mechanism similar to vouching.

Trust by association

Trust by association (TbyA) [7] has been designed for deployment in a peer-to-peer environment. It utilizes invitations to add new entities to the network and links the reputation values of the *inviter* and *invitee*, similar to vouching. More specifically, inviters may be punished for the bad behaviour of their invitees, while incentivized by profiting from the good reputation of their invitees and the rewards for growing the network. Due to these reputational incentives, it is assumed that users will only invite people they already have experience with from another channel, i.e. the acquaintance. TbyA assumes the following properties of the network:

- *Invitation-only network* – entities can only join the network through invitation.
- *Homogeneous Resource or Service* – entities participate in the network for a common type of resource or service.
- *Bounded Existing Reputation Mechanism* – there exists an underlying reputation mechanism, such that the resulting reputation values are bounded within a fixed interval. Kellett et al. [7] suggests the usage of EigenTrust [29].
- *Central Point of Calculation* – there exists a central machine on which all calculations take place.

In its simplest form, the reputation value R for entity i, R_i, is defined by:

$$R_i = (1 - \alpha)U_i + \alpha \frac{\sum_{j \in N_i} U_j}{|N_i|}$$

where U_i is a function returning the reputation of entity i according to the underlying reputation mechanism, N_i refers to the set of invitees invited by entity i, and α is some value $0 \leq \alpha < 1$ and is assumed to be 0 when $|N_i| = 0$. While [7] only uses this formula as a starting point to introduce enhancements, the general idea remains unchanged. These enhancements include rewarding network growth by varying α depending on the amount of entities invited and support for recursive reputation, i.e. the reputation of the invitee's of entity i's invitees affects entity.

In an effort to measure TbyA's efficacy, a simulation was performed. It was found that TbyA performs well in the case where there exists an external party capable of identifying malicious entities and punishing their inviters. TbyA is said to be able to turn lawless peer-to-peer networks into networks of benevolent peers, but requires future work on decentralized methods of identifying malicious entities.

TbyA uses elements we have previously seen in Souche similar to vouching, as it punishes the inviter for any bad behaviour shown by their invitees. However, besides social reputation mechanisms in which you need a voucher to participate, there also exist less strict mechanisms. One such mechanisms is SocialTrust, in which anyone can participate, but where existing social ties are useful. SocialTrust uses the notion that friends are more trusted than strangers [55].

SocialTrust

SocialTrust [8] attempts to combine entity reputation values as well as friendships to provide the best QoS in a decentralized network. SocialTrust's main goal is to attempt being served by a friend or, if no friend is available, the server with the highest global reputation, provided by a *trusted authority*. First of all, SocialTrust defines two types friendship, namely 'friends' and 'partners', both being bidirectional relationships. An entity can choose their own friends based on their experiences in the offline physical world and send a 'friend request'. However, partners are assigned by the *trusted authority* and are defined as entities with whom a certain entity has had many interactions with. In order to participate in a partnership, both entities must have a reputation larger than a certain partner threshold.

When some entity i requires a certain service or resource, it first composes a list of all possible entities which may pose as server. In this process, the reputation mechanism takes the current load of entities into account, such that overloaded entities are not included in the list of possible servers. After composing the list, entity i scans for any friends or partners and, if present, selects one of these to request the service

or resource. If such friend or partner does not exist, entity i queries the trusted authority for the reputations of the possible servers and chooses the server with the highest reputation.

In SocialTrust, each entity is assigned an *impact factor*, which represents both their reputation the amount of damage they could inflict and is used to calculate entity's new reputation after an interaction, depending on whether it cooperates. The impact factor T is defined such that:

$$T(i) = \beta \frac{R(i)}{R_{max}} + (1 - \beta) \frac{D(i)}{D_{max}}$$

where R is a function returning an entity's reputation, R_{max} is the maximum achievable reputation, $D(i)$ represents the number of friends and partners, D_{max} is the maximum number of friends and partners and β is some value $0 \leq \beta \leq 1$. After each interaction, the client will provide a service rating of the server, which helps the trusted authority to calculate the new reputations by taking into account the impact factor.

We consider two cases: an interaction in which both the server and client are cooperative and an interaction in with the server is cooperative, but the client is non-cooperative. In the first case, the client will, subsequently to the interaction, send a service rating to the trusted authority, in which it rates the server with some value Y, such that $0 < Y \leq 1$. A cooperative server will accept this rating and the servers reputation increases by $\alpha(1 + T_c \cdot Y)$, where T_c is the client's impact factor and $0 \leq \alpha \leq 1$. The client's reputation will increase with α. On the other hand, we consider the cases where the client is non-cooperative and provides no feedback or negative feedback, while the server provided honest work (the trusted authority checks this by verifying the signatures on the request and response). In this case, the server is assigned α reputation and the client loses $-\alpha(1+T_c)$ reputation. Similar reputation assignments are presented in [8] for non-cooperative servers, in which the server loses reputation. Note that the more reputation and friends/partners an entity has, the more their reputation is affected in interactions, promoting honest work for all, regardless of reputation.

In a performance evaluation, SocialTrust has shown stronger capabilities in excluding non-cooperative entities from the network compared to EigenTrust [29], as well as obtaining a more accurate model mapping an entity's reputation to its benevolence.

Acquaintance-based social reputation mechanism using concepts like vouching often offer built-in defenses against attacks. However, bootstrapping such mechanisms is a challenge, as they often require an initial set of trusted entities from which all remaining participants join the network. The concept of *implicit vouching* as introduced by Souche might open the opportunities for deploying vouching-based mechanisms, but may inadvertently punish innocent entities. Reputation mechanisms such as SocialTrust suffer less from the bootstrap problem, but have weaker defences for filtering malicious entities.

6 Neighbourhoods

The final scope is focused on the notion of neighbourhoods. In a social context, one's neighbourhood often determines their opportunities and success in later stages of life [56]. Moreover, social groups often arise from these neighbourhoods. These groups may determine one's reputation as it has been shown that social groups are often assigned a single reputational value [57].

Similar concepts have been applied in the design of reputation mechanisms. One such reputation mechanism is GroupRep, in which entity j's reputation in entity i's perspective may be determined by their group if no direct interaction has occurred.

GroupRep

Based on the assumption that in large peer-to-peer networks, two peers will not often interact more than once, making it hard to profit from direct experiences between peers, GroupRep [9] adopts the notion of groups to calculate reputational values. By assuming that users with similar interests in a peer-to-peer environment have constructed virtual groups, GroupRep provides a framework for calculating reputational values between groups, between groups and peers and between peers.

In GroupRep, the notion of a trust graph is applied on two scales. On the first scale, every node in the trust graph are groups of entities in which the edges represent reputations from the group perspective. The second scale considers all nodes individual entities, in which the edges represent reputation values based on direct experiences between entities. Moreover, GroupRep defines utility u and costs c, which represent the gain and costs from interactions with other entities or entity groups. In general, reputation is calculated by $\frac{c_{ij}-u_{ij}}{c_{ij}+u_{ij}}$, where c_{ij} represents the cumulative cost some entity or group j has brought entity or group i and u_{ij} represents the cumulative utility. However, if $c_{ij} + u_{ij} = 0$, a fall-back policy is applied in which a path (on the group-based trust graph) is searched between $G(i)$ and $G(j)$, where G is a function returning an entity's group. Note that for all groups along this path, including $G(i)$, the most trusted group is selected for each next step. The reputation of this path is equivalent to the minimum reputation edge on the path. However, if no such path exists, a stranger policy is applied, in which the reputation is calculated using the cumulative utility and cost for all previous interactions with strangers. Note that GroupRep will always first attempt to find direct reputation values on the trust graph on entity-level, however, if no such direct edge exists, the group reputation is used for determining a reputation value. After an interaction, entity i updates its local information, creating an edge in the entity

trust graph, and sends the rating to its group $G(i)$, which then may sends the rating to group $G(j)$.

Furthermore, GroupRep introduces a methodology for detecting malicious entities through clustering entities within groups. By assuming two entities as similar when they have similar reputations on the entities they both have had interactions with, clustering can take place. It is assumed that a maximum cluster of similar entities will take shape, in which all entities are deemed credible.

GroupRep has been compared against two existing reputation mechanisms on the performance against malicious collusive attacks. It was shown that GroupRep achieves a higher ratio of success queries (ratio of peers satisfied with the result of the interaction) and a higher satisfaction level, where satisfaction represents the average ratio of cumulative authentic file sizes to cumulative inauthentic file sizes. However, the scope of this evaluation was somewhat limited and did not include comparison against any well-known reputation mechanisms.

While entities are still somewhat free to choose which group to join when using GroupRep, there also exist more discriminative approaches, which may be associated with originative discrimination. Such methodologies are commonly adopted in email spam measures where IP addresses are blacklisted. One such mechanism is IPGroupRep (name similarity with GroupRep is coincidental), which aggressively groups IP addresses into blocks based on subnets and assigns single reputation values to these groups based on their behaviour.

IPGroupRep

IPGroupRep [10] is an aggressive reputation mechanism for calculating a reputation for IP blocks based on existing spam classifiers. It only considers groups of IP addresses, rather than leveraging individual reputations with a group reputation. In [10], it is suggested to consider cluster IP into blocks of 256 by naively assuming the first 24 bits of all IP addresses in a block static, similar to a 255.255.255.0 subnet mask. An IP block's reputation should be decreased when a spam message originating from this group is detected, while it should be increased upon sending legitimate messages. Note that IPGroupRep is in itself not capable or designed to detect spam, but rather to combine the outputs of several existing spam detection mechanisms and combine these into a single reputation value.

For each group, a sum r and s are defined, representing the aggregation of positive and negative spam feedback respectively, provided by the numerous spam detection mechanisms. IPGroupRep applies a beta distribution, where $\alpha = r + 1$ and $\beta = s + 1$ and assumes the expected value E(p) to be the reputation value, such that:

$$E(p) = \frac{r+1}{r+s+2}$$

If this value $E(p)$ is larger than some threshold $T_{threshold}$, the group can be assumed trustworthy.

In evaluation it was found that this reputation mechanism shows very high precision and accuracy compared to existing reputation mechanisms used for the protection of mail servers. However, we argue that this method may negatively affect innocent parties within a group by disregarding the individual reputations. A possible solution to alleviate this is by decreasing the group sizes or automatically detect dynamic IP address blocks which may be used for spam [20].

While the usage of groups may be effective against spamming and the danger of strangers, it is very generative and should be implemented cautiously such that malicious entities cannot hide in highly reputed groups and enjoy their benefits.

7 Conclusion

In this paper, we have discussed numerous social phenomena on different scales and reviewed social reputation mechanisms directly adopting the social phenomena as core component. First, we focused on the individual scope, in which every entity is responsible for their own reputation and entities may refer to each other based on past interactions, increasing each other's reputation by performing honest work. Secondly, we reviewed the acquaintances scope, where mechanisms may benefit from existing social ties to create more secure environments through vouching and friends. In this space, the existing trust relations are essential and may heavily influence one's reputation, compared to the individual scope. Finally, we reviewed mechanisms in the neighbourhood scope, in which entities may be part of a group which can greatly affect their reputation. Over the years, many reputation mechanisms have been proposed, evaluated and criticised. However, the holy grail of a social reputation mechanism creating secure online trust is yet to be invented.

References

[1] P. Van Aelst, F. Toth, L. Castro, V. Štětka, C. d. Vreese, T. Aalberg, A. S. Cardenal, N. Corbu, F. Esser, D. N. Hopmann, *et al.*, "Does a crisis change news habits? a comparative study of the effects of covid-19 on news media use in 17 european countries," *Digital Journalism*, vol. 9, no. 9, pp. 1208–1238, 2021.

[2] A. M. Enders, J. E. Uscinski, M. I. Seelig, C. A. Klofstad, S. Wuchty, J. R. Funchion, M. N. Murthi, K. Premaratne, and J. Stoler, "The relationship between social media use and beliefs in conspiracy theories and misinformation," *Political behavior*, pp. 1–24, 2021.

[3] M. Wang, F. Tao, Y. Zhang, and G. Li, "An adaptive and robust reputation mechanism for p2p network," in *2010 IEEE International Conference on Communications*, pp. 1–5, 2010.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[5] B. Nasrulin, G. Ishmaev, and J. Pouwelse, "Meritrank: Sybil tolerant reputation for merit-based tokenomics," *arXiv preprint arXiv:2207.09950*, 2022.

[6] Y. Xie, F. Yu, Q. Ke, M. Abadi, E. Gillum, K. Vitaldevaria, J. Walter, J. Huang, and Z. M. Mao, "Innocent by association: early recognition of legitimate users," in *Proceedings of the 2012 ACM conference on Computer and communications security*, pp. 353–364, 2012.

[7] M. Kellett, T. Tran, and M. Li, "Trust by association: A meta-reputation system for peer-to-peer networks," *Computational Intelligence*, vol. 27, no. 3, pp. 363–392, 2011.

[8] K. Chen, H. Shen, K. Sapra, and G. Liu, "A social network based reputation system for cooperative p2p file sharing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 8, pp. 2140–2153, 2015.

[9] H. Tian, S. Zou, W. Wang, and S. Cheng, "A group based reputation system for p2p networks," in *International Conference on Autonomic and Trusted Computing*, pp. 342–351, Springer, 2006.

[10] H. Zhang, H. Duan, W. Liu, and J. Wu, "Ipgrouprep: A novel reputation based system for anti-spam," in *2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, pp. 513–518, 2009.

[11] W. Zhang, S. Zhang, and S. Guo, "A pagerank-based reputation model for personalised manufacturing service recommendation," *Enterprise Information Systems*, vol. 11, no. 5, pp. 672–693, 2017.

[12] J. Bi, J. Wu, and W. Zhang, "A trust and reputation based anti-spim method," in *IEEE INFOCOM 2008-The 27th Conference on Computer Communications*, pp. 2485–2493, IEEE, 2008.

[13] J. Wang, J. Peng, and D. Zhang, "Research on dynamic reputation management model based on pagerank," in *2008 International Conference on Computer Science and Software Engineering*, vol. 3, pp. 814–817, IEEE, 2008.

[14] J. M. Pujol, R. Sangüesa, and J. Delgado, "Extracting reputation in multi agent systems by means of social network topology," in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pp. 467–474, 2002.

[15] J. Sabater and C. Sierra, "Regret: reputation in gregarious societies," in *Proceedings of the fifth international conference on Autonomous agents*, pp. 194–195, 2001.

[16] X. Wu, J. He, and F. Xu, "A group-based reputation mechanism for mobile p2p networks," in *International Conference on Grid and Pervasive Computing*, pp. 410–421, Springer, 2009.

[17] L. Sun, L. Jiao, Y. Wang, S. Cheng, and W. Wang, "An adaptive group-based reputation system in peer-to-peer networks," in *International Workshop on Internet and Network Economics*, pp. 651–659, Springer, 2005.

[18] W. Ji, S. Yang, D. Wei, and W. Lu, "Garm: A group - anonymity reputation model in peer-to-peer system," in *Sixth International Conference on Grid and Cooperative Computing (GCC 2007)*, pp. 481–488, 2007.

[19] D. He, Z. Peng, L. Hong, and Y. Zhang, "A social reputation management for web communities," in *International Conference on Web-Age Information Management*, pp. 167–174, Springer, 2011.

[20] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, and T. Wobber, "How dynamic are ip addresses?," in *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications*, pp. 301–312, 2007.

[21] A. Thomas, "Rapid: Reputation based approach for improving intrusion detection effectiveness," in *2010 Sixth International Conference on Information Assurance and Security*, pp. 118–124, IEEE, 2010.

[22] H. Esquivel, A. Akella, and T. Mori, "On the effectiveness of ip reputation for spam filtering," in *2010 Second International Conference on COMmunication Systems and NETworks (COMSNETS 2010)*, pp. 1–10, 2010.

[23] L. Xiong and L. Liu, "Peertrust: supporting reputation-based trust for peer-to-peer electronic communities," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 7, pp. 843–857, 2004.

[24] P. Gauthier, B. Bershad, and S. D. Gribble, "Dealing with cheaters in anonymous peer-to-peer networks," *Proceedings of technical report of University of Washington*, 2004.

[25] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt, "Certified reputation: How an agent can trust a stranger," AAMAS '06, (New York, NY, USA), p. 1217–1224, Association for Computing Machinery, 2006.

[26] A. Abdul-Rahman and S. Hailes, "Supporting trust in virtual communities," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, pp. 9 pp. vol.1–, 2000.

[27] S. N. L. C. Keung and N. Griffiths, "Using recency and relevance to assess trust and reputation," in *Proceedings of AISB 2008 Symposium on Behaviour Regulation in Multi-Agent Systems*, vol. 4, pp. 13–18, 2008.

[28] M. Rogers and S. Bhatti, "How to disappear completely: A survey of private peer-to-peer networks," *networks*, vol. 13, p. 14, 2007.

[29] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The eigentrust algorithm for reputation management in p2p networks," in *Proceedings of the 12th international conference on World Wide Web*, pp. 640–651, 2003.

[30] Y. Gil and V. Ratnakar, "Trusting information sources one citizen at a time," in *International Semantic Web Conference*, pp. 162–176, Springer, 2002.

[31] J. R. Douceur, "The sybil attack," in *International workshop on peer-to-peer systems*, pp. 251–260, Springer, 2002.

[32] L. Mui, M. Mohtashemi, and A. Halberstadt, "A computational model of trust and reputation," in *Proceedings of the 35th annual Hawaii international conference on system sciences*, pp. 2431–2439, IEEE, 2002.

[33] A. Binks, "The art of phishing: past, present and future," *Computer Fraud & Security*, vol. 2019, no. 4, pp. 9–11, 2019.

[34] B. Bogenschneider, "ebay frauds: Specific illustrations and analysis," *Loyola Consumer Law Review, Forthcoming*, 2021.

[35] S. Read, "Google fined €220m in france over advertising abuse," *BBC News*, Jun 2021.

[36] S. Ghose, "The red queen and the inevitability of the amazoogle business model," May 2018.

[37] J. A. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D. H. Epema, M. Reinders, M. R. Van Steen, and H. J. Sips, "Tribler: a social-based peer-to-peer system," *Concurrency and computation: Practice and experience*, vol. 20, no. 2, pp. 127–138, 2008.

[38] G. Swamynathan, K. C. Almeroth, and B. Y. Zhao, "The design of a reliable reputation system," *Electronic Commerce Research*, vol. 10, no. 3, pp. 239–270, 2010.

[39] D. E. Saputra, "Defining trust in computation," in *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*, pp. 161–166, IEEE, 2020.

[40] B. N. Levine, C. Shields, and N. B. Margolin, "A survey of solutions to the sybil attack," *University of Massachusetts Amherst, Amherst, MA*, vol. 7, p. 224, 2006.

[41] S. Seuken and D. C. Parkes, "On the sybil-proofness of accounting mechanisms," 2011.

[42] J. Sabater and C. Sierra, "Review on computational trust and reputation models," *Artificial intelligence review*, vol. 24, no. 1, pp. 33–60, 2005.

[43] M. Feinberg, R. Willer, and M. Schultz, "Gossip and ostracism promote cooperation in groups," *Psychological science*, vol. 25, no. 3, pp. 656–664, 2014.

[44] T. Hogg and L. Adamic, "Enhancing reputation mechanisms via online social networks," in *Proceedings of the 5th ACM Conference on Electronic Commerce*, pp. 236–237, 2004.

[45] F. A. Massucci and D. Docampo, "Measuring the academic reputation through citation networks via pagerank," *Journal of Informetrics*, vol. 13, no. 1, pp. 185–201, 2019.

[46] A. Cheng and E. Friedman, "Manipulability of pagerank under sybil strategies," 2006.

[47] T. T. A. Dinh and M. Ryan, "A sybil-resilient reputation metric for p2p applications," in *2008 International Symposium on Applications and the Internet*, pp. 193–196, IEEE, 2008.

[48] G. Danezis and S. Schiffner, "On network formation,(sybil attacks and reputation systems)," in *DIMACS Workshop on Information Security*

Economics, pp. 18–19, 2006.

[49] W. Chang and J. Wu, "A survey of sybil attacks in networks," *Sensor Networks for Sustainable Development*, pp. 497–533, 2012.

[50] A. Stannat, C. U. Ileri, D. Gijswijt, and J. Pouwelse, "Achieving sybil-proofness in distributed work systems.," in *AAMAS*, pp. 1263–1271, 2021.

[51] "European digital identity," Oct 2022.

[52] "Vouch," in *Cambride Dictionary*, Cambridge University Press.

[53] "How to accept a vouch as evidence of someone's identity," Oct 2020.

[54] D. Lauterbach, H. Truong, T. Shah, and L. Adamic, "Surfing a web of trust: Reputation and reciprocity on couchsurfing.com," in *2009 International Conference on Computational Science and Engineering*, vol. 4, pp. 346–353, 2009.

[55] D. Li, L. Eden, M. A. Hitt, and R. D. Ireland, "Friends, acquaintances, or strangers? partner selection in r&d alliances," *Academy of management journal*, vol. 51, no. 2, pp. 315–334, 2008.

[56] B. S. Graham, "Identifying and estimating neighborhood effects," *Journal of Economic Literature*, vol. 56, pp. 450–500, June 2018.

[57] N. Masuda, "Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation," *Journal of Theoretical Biology*, vol. 311, pp. 8–18, 2012.