Achieving Sybil-Proofness in Distributed Work Systems

Alexander Stannat
TU Delft
Delft, The Netherlands
a.w.stannat@tudelft.nl

Can Umut Ileri
TU Delft
Delft, The Netherlands
c.u.ileri@tudelft.nl

Dion Gijswijt
TU Delft
Delft, The Netherlands
d.c.gijswijt@tudelft.nl

Johan Pouwelse
TU Delft
Delft, The Netherlands
j.a.pouwelse@tudelft.nl

ABSTRACT

In a multi-agent system where agents provide quantifiable work for each other on a voluntary basis, reputation mechanisms are incorporated to induce cooperation. Hereby agents assign their peers numerical scores based on their reported transaction histories. In such systems, adversaries can launch an attack by creating fake identities called Sybils, who report counterfeit transactions among one another, with the aim of increasing their own scores in the eyes of others. This paper provides new results about the Sybil-proofness of reputation mechanisms. We revisit the impossibility result of Seuken and Parkes (2011), who show that strongly-beneficial Sybil attacks cannot be prevented on reputation mechanisms satisfying three particular requirements. We prove that, under a more rigorous set of definitions of Sybil attack benefit, this result no longer holds. We characterise properties under which reputation mechanisms are susceptible to strongly-beneficial Sybil attacks. Building on our results, we propose a minimal set of requirements for reputation mechanisms to achieve resistance to such attacks, which are stronger than the results by Cheng and Friedman (2005), who show Sybil-proofness of certain asymmetric reputation mechanisms.

KEYWORDS

Sybil Attacks; Cooperation; Reputation; Impossibility Results

ACM Reference Format:
Alexander Stannat, Can Umut Ileri, Dion Gijswijt, and Johan Pouwelse. 2021. Achieving Sybil-Proofness in Distributed Work Systems. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 9 pages.

1 INTRODUCTION

Multi-agent work systems such as P2P file sharing networks rely on the cooperation of participants to function effectively, whereby agents perform work for one another voluntarily without a central orchestrator. Cooperation can be enforced in such systems through evolutionary mechanisms of indirect reciprocity, observed in biological communities, which are based on reputation schemes that help agents in deciding who to interact with and who to shun.

However, in decentralised networks that do not restrict the involvement of new users, reputation mechanisms can be manipulated by Sybil attacks, in which a malicious agent creates multiple fake identities who report fraudulent transactions about one another to honest agents with the goal to artificially increase their reputation with the network, thereby convincing the remaining honest agents to perform more work for them than they would, had the reputation not been fraudulently increased.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

There have been many studies towards understanding the dynamics of Sybil attacks on reputation mechanisms [5, 10–13]. Seuken and Parkes introduce a mathematical framework of algorithms with which an agent can assign all of its peers some score reflecting their level of cooperativeness in the system, which they refer to as accounting mechanisms [11]. For instance, an agent i may assign another agent j a score given by j's overall net contribution to the network, i.e., aggregated contribution minus consumption of work. Given this reputation mechanism, a malicious j may create many Sybil identities that falsely report having received work from j.

This example highlights the challenge of Sybil-proofing reputation mechanisms, which lies in an inherent trade-off between a reputation mechanism's ability to induce cooperation and its resistance to Sybil attacks. A reputation mechanism successfully induces cooperation if contributions increase an agent's reputation while the consumption of resources reduces it, thereby rewarding altruism and penalising free riding. However, contributions made must be weighted differently based on the parties involved as the authenticity of interactions is not guaranteed, which often comes at the expense of cooperation inducivenes. In this paper we focus primarily on the Sybil-proofness aspect of reputation mechanisms, disregarding any theoretical properties reputation mechanisms may require to induce cooperation. Note that there are trivial, but useless schemes satisfying our requirements on Sybil-proofness and we assume any reasonable reputation mechanism will successfully induce cooperation, thereby excluding these types of mechanisms.

Seuken and Parkes proved that any reputation mechanism satisfying three requirements will be susceptible to Sybil attacks through which the attacker can gain infinitely more work than they have performed [11]. They prove this by constructing a particular Sybil attack on a given graph and concluding that while the amount of work the attacker gained was finite, the amount of work that was performed was zero. While this was a seminal and consequential result to obtain, we argue that it entails an inaccuracy that can be attributed to a lack of rigour in their definitions of cost and profit of Sybil attacks. Our contributions to the state-of-the-art are threefold.

- We define for the first time the *cost* and *profit* of a Sybil attack in Section 4 and show that under these definitions the impossibility result of [11] no longer holds, in Section 5 [1].
- In Section 5, we introduce a pair of novel requirements, called *parallel-report responsiveness* and *serial-report responsiveness*, with which we obtain a new pair of impossibility results, derived from those of [11].
- We then invert the intuition behind the results above and introduce a set of requirements for reputation mechanisms to achieve Sybil-Proofness in Section 6.

[1] Upon a correspondence with the authors, they agreed with our findings.

2 RELATED WORK

Most previous work on reputation mechanisms in multi-agent systems has focused on the domain of P2P file sharing. BitTorrent [1] uses a mechanism called tit-for-tat to punish free riding agents with the help of a choking algorithm. Despite being an effective game-theoretical strategy in enforcing cooperation, tit-for-tat relies on repeated encounters between agents which are not guaranteed in large networks. The distributed reputation mechanism in [16] is based on the PageRank algorithm [9], which disincentivises free riding better than tit-for-tat, but has been shown to be susceptible to basic types of Sybil attacks [3]. Distinguishing between trust and reputation, EigenTrust [4] computes the global reputation of peers through local trust values assigned to them by others. As it relies on pre-trusted agents, it cannot be regarded as fully distributed. BarterCast [6] is based on the maximum flow between two nodes in the interaction graph and is shown to be efficient in punishing free riders, but is susceptible to a type of Sybil attack known as a parallel attack. Netflow [7] mitigates the effects of Sybil attacks on BarterCast by introducing an additional constraint of vertex capacities, however it is ineffective at discerning between cooperators and free riders. Seuken and Parkes [11] introduce a set of impossibility results in which they state that any accounting mechanism satisfying a transitive property is susceptible to Sybil attacks. Cheng and Friedman [2] show an impossibility result for symmetric reputation mechanisms and provide a set of requirements reputation mechanisms should satisfy to be Sybil-proof. Note that their impossibility result is based on a global reputation mechanism, while ours is based on personalised reputation and a different definition of symmetry, given a seed node. Secondly, our definitions of Sybil-proofness are stricter than those in [2] and our final results on Sybil-proofness are therefore stronger than theirs.

3 PRELIMINARIES

We consider a network of distributed agents denoted by a directed graph $G = (V, E, w)$, called the work graph, where V is the set of agents and E is the set of edges between agents. An edge $(i, j) \in E$ represents work performed by j and consumed by i. The function $w : V \times V \to \mathbb{R}_{\geq 0}$ denotes the weight of the edges, i.e., $w(i, j)$ represents the total amount of work performed by j for i.

Agents do not have full knowledge of G, but keep track of their interactions with other agents. A *history* of an agent i is denoted by H_i and includes all edges $(k, j) \in E$ where either $i = k$ or $i = j$. Agents share their history with others using a gossip protocol, whereby agents query one another about their respective transaction histories, which they consequently exchange mutually. The honest reporting of transaction histories can be achieved through a mechanism known as Drop-Edge, introduced in [13] or alternatively through digital signatures in a distributed data structure, such as TrustChain [8]. We do not make any assumptions on how honest reporting is achieved and assume it is ensured by a mechanism that is outside the scope of this paper. We denote the reported overall work performed by agent k for j as reported to i by j by $w_i^j(j, k)$, which is set to zero, if i does not receive any report.

Using this information, each agent i can construct a *subjective work graph* $G_i = (V_i, E_i, w_i)$. For integrity of notation, we call G the *objective work graph*.

3.1 Reputation Mechanism

Agents use a reputation mechanism R to assign numerical scores to other agents, usually based on some graph theoretical impact measure applied to the subjective work graph. We note that Seuken et al. [10] prefer to use the term *accounting mechanism* instead of reputation mechanism, arguing that 1) accounting mechanisms use averaging rather than aggregation of historical values, and 2) work performed by an agent i for j does not affect the reputation of j negatively. We disagree with both, since 1) there exists previous work on reputation aggregation, and 2) there exists previous work on reputation mechanisms where a contribution by i to j negatively affects the reputation of j [6]. Therefore, we choose to use the general term reputation mechanism, while not excluding the aggregation principle and the negative effect of consumption on the reputation.

Definition 3.1 (Reputation Score)**.** A reputation score $R_i(G_i, j)$ is a value assigned by i to an agent j, given a subjective work graph G_i, using some underlying algorithm. Literally,

$$R_i(G_i, j) \in \mathbb{R} \quad \forall j \in V_i \backslash \{i\}.$$

Every agent i then obtains a set of *reputation scores* for all agents in its subjective work graph, excluding itself, denoted as

$$R_i(G_i) := \{R_i(G_i, j) \mid j \in V_i \backslash \{i\}\}.$$

An agent that is willing to perform some work is queried by agents that want to receive some work and is able to choose who to perform work for from this set. The set of all agents that query i at a particular point in time is called a *choice set* and denoted $C_i \subset V_i \backslash \{i\}$. It can be of variable size or even empty. Given the subjective work graph G_i, the choice set C_i and a reputation mechanism R_i, an *allocation policy* A_i returns a subset of the choice set as the peers to receive work from i. Formally, A_i is defined as:

$$A_i : \mathbb{R}^{|V_i|-1} \times \mathcal{P}(V) \to \mathcal{P}(V) \text{ with } A_i(R_i(G_i), C_i) \subset C_i.$$

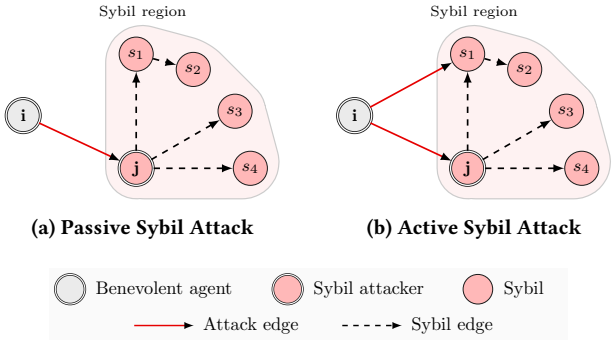
As an example one may choose the *winner-takes-all allocation policy* which outputs a single agent with the highest reputation in the choice set, breaking ties randomly when necessary.

We assume without loss of generality that, whenever an agent chooses to perform work for a peer, they perform the same amount of work. However, our findings apply to non-standard amounts of work as well.

3.2 Sybil Attacks

We assume two types of agents; namely honest and Sybil. *Honest agents* control only one identity, while Sybils are the agents created by an attacker in a Sybil attack. While one may also differentiate between Sybil identities and the attacker that creates them, the network cannot distinguish between these identities and we therefore treat the agent launching the attack as just another Sybil.

Every Sybil attack entails real transactions in which the attacker makes a legitimate donation to honest agents either directly or through its Sybils. In the work graph these legitimate transactions are called attack edges. In Definition 6.4 we provide a requirement that necessitates the existence of attack edges in Sybil attacks.



(a) Passive Sybil Attack (b) Active Sybil Attack

Figure 1: Examples of passive and active Sybil attacks

Definition 3.2 (Sybil Attack)**.** Given an objective work graph $G = (V, E, w)$, an attacker j performs a Sybil attack σ_S by introducing the following elements to the work graph:

- A set of Sybil identities $S = \{j, s_1, \ldots, s_m\}$, each of which is called a **Sybil** and is indistinguishable from an honest agent by other agents,
- a set of **Sybil edges** $E_S \subset S \times S$ with edge weights $w_S : S \times S \to \mathbb{R}_{\geq 0}$,
- a set of **attack edges** E_a with weights $w_a : V \times S \to \mathbb{R}_{\geq 0}$.

The Sybil attack alters the work graph and we obtain the new graph G' as follows:

$$G' := G \downarrow \sigma_S = (V', E', w') = (V \cup S, E \cup E_S \cup E_a, w')$$

where

$$w'(u,v) = \begin{cases} w(u,v), & \text{if} \quad u,v \in V \\ w_S(u,v), & \text{if} \quad u,v \in S \\ w_a(u,v), & \text{if} \quad u \in V, v \in S \end{cases}.$$

We define a Sybil attack on a subjective work graph equivalently by $G'_i := G_i \downarrow \sigma_S$.

We differentiate Sybil attacks based on the plurality of agents connected to attack edges. Considering a subjective work graph, in a *passive Sybil attack*, attack edges are only connected to one and the same agent, which we assume w.l.o.g. is the attacker j, i.e., $w'(k,s) = 0$ for all $s \in S \backslash \{j\}, k \in V$. In an *active Sybil attack*, however, every Sybil may be connected to the honest region of the network. Active and passive attacks are visualised in Figure 1.

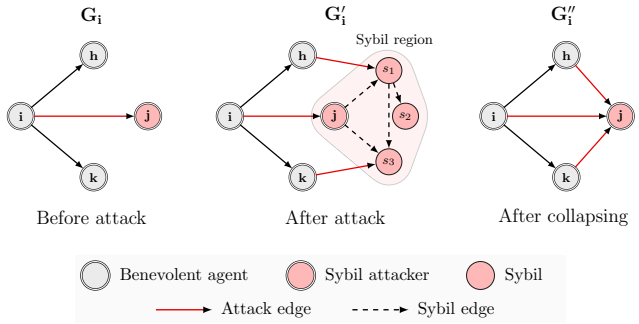


Figure 2: Example vertex identification in a Sybil region.

4 COST AND PROFIT OF SYBIL ATTACKS

We now define the benefit of a Sybil attack given by its cost and its profit. We consider the evolution of the network over time, whereby we assume the network evolves chaotically over time. At each point in time agents query one another and decide which of the agents in their choice set should receive some work. We assume a sequence of work graphs $(G^{(t)})_{t \geq 0}$, where we assume $G^{(0)}$ is the work graph at the time of the attack. The profit of a Sybil attack is the amount of work the attacker can consume as a result of their attack given by

$$\omega^+_{\text{work}}(\sigma_S) := \lim_{t \to \infty} \sum_{i \in V^{(t)}} \sum_{s \in S} w^{(t)}(s,i).$$

Analogously, we can define the cost of a Sybil attack as the amount of work the attacker has had to invest into their attack.

$$\omega^-_{\text{work}}(\sigma_S) := \sum_{i \in V} \sum_{s \in S} w'(i,s).$$

While the value for the cost of a Sybil attack is easy to determine, the profit of a Sybil attack is based on a prediction of how the network evolves over time and how much work the attacker can consume. No agent can predict the behavior of other agents in the network and therefore it is not clear who will decide to serve the attacker in the future. Consequently, the profit of a Sybil attack is impractical to compute in any generic setting. We therefore introduce a pair of new definitions for the cost and profit of a Sybil attack in terms of reputation scores, which will serve as proxies for the definitions above and are useful as they are more straightforward to determine. The profit of a Sybil attack is given by

$$\omega^+_{\text{rep}}(\sigma_S) := \sum_{i \in V} \sum_{s \in S} R_i(G'_i, s).$$

With the definition of $\omega^-_{\text{work}}(\sigma_S)$, we capture the amount of work invested into the network by a Sybil attacker, i.e., the aggregated weight of the attack edges. We can define its analogue in terms of reputation as the aggregated scores a Sybil attacker has *earned* through their honest work. All edges that do not enter or leave the Sybil region should therefore be disregarded and any increase in reputation that the Sybil attackers may gain through the Sybil-internal edges should not be taken into account.

The cost of a Sybil attack in terms of reputation scores is therefore determined by performing an operation of vertex identification on S yielding a new work graph G'' with a new agent $\mathbf{s}$, as visualised in Figure 2. In formula it is given by

$$\omega^-_{\text{rep}}(\sigma_S) := \sum_{i \in V''} R_i(G''_i, \mathbf{s}).$$

With these rigorous definitions for the cost and profit, we can define the benefit of a Sybil attack. We say that a Sybil attack is

- **strongly beneficial** if $\omega^+(\sigma_S) > 0$ and $\omega^-(\sigma_S) = 0$ or if $\lim_{|S| \to \infty} \frac{\omega^+(\sigma_S)}{\omega^-(\sigma_S)} = \infty$,
- **weakly beneficial** if $\omega^+(\sigma_S) > 0$ and $\omega^-(\sigma_S) > 0$ and $\exists c > 1 : \lim_{|S| \to \infty} \frac{\omega^+(\sigma_S)}{\omega^-(\sigma_S)} \leq c$.

Inversely, we say that a reputation mechanism is resistant to strongly (weakly) beneficial Sybil attacks if the conditions above are not satisfied for any arbitrarily large Sybil attack. In particular, if the conditions above are satisfied for $\omega^{\pm}_{\text{work}}(\sigma_S)$ we say beneficial/resistant in terms of work and otherwise we say in terms of reputation scores.

In this work, we focus on resistance to **strongly** beneficial attacks in which the ratio of the cost and profit is finite. One may argue that this is a rather loose requirement, however we point out that a finite ratio of cost and profit ensures that an attacker who wants to consume infinite work from the network must also contribute infinite work. That way no Sybil attack can compromise the overall operability of the network, even if scaled to infinite size.

Note that these definitions are stronger than value and rank Sybil-proofness introduced by Cheng and Friedman [2]. A reputation mechanism is value Sybil-proof if a Sybil node can obtain a higher reputation than the attacker, and rank Sybil-proof if it can achieve a higher reputation than a node that previously had a higher reputation than the attacker. This is not equivalent to strongly/weakly beneficial Sybil attacks, where we evaluate the sum of the Sybils' reputation scores. A Sybil attack where every Sybil gains some reputation that is less than that of i can still be strongly beneficial. The maxflow mechanism, for instance, is rank and value Sybil-proof, as no Sybil node can obtain reputation scores larger than i, but it is susceptible to strongly beneficial Sybil attacks.

While it is our goal for reputation mechanisms to be resistant to strongly beneficial Sybil attacks in terms of work, it is impossible to determine the profit of a Sybil attack in terms of work, due to the randomness of network interactions. We therefore use the proxy of reputation scores. We now turn our attention to ensuring the effectiveness of this proxy by introducing a requirement such that a Sybil attack that is strongly beneficial in terms of work is also strongly beneficial in terms of reputation scores, and vice versa.

Definition 4.1 (Representative)**.** We say a reputation mechanism R is *weakly representative* if it holds for any work graph $G = (V, E, w)$ and any Sybil attack σ_S that

$$\lim_{|S|\to\infty} \frac{\omega^{+}_{\text{rep}}(\sigma_S)}{\omega^{-}_{\text{rep}}(\sigma_S)} < \infty \implies \lim_{|S|\to\infty} \frac{\omega^{+}_{\text{work}}(\sigma_S)}{\omega^{-}_{\text{work}}(\sigma_S)} < \infty.$$

Subsequently, we call a reputation mechanism R *strongly representative* if it holds that

$$\lim_{|S|\to\infty} \frac{\omega^{+}_{\text{rep}}(\sigma_S)}{\omega^{-}_{\text{rep}}(\sigma_S)} < \infty \iff \lim_{|S|\to\infty} \frac{\omega^{+}_{\text{work}}(\sigma_S)}{\omega^{-}_{\text{work}}(\sigma_S)} < \infty.$$

If a reputation mechanism is resistant to strongly beneficial Sybil attacks in terms of reputation, and weakly representative, then it is also resistant to strongly beneficial Sybil attacks in terms of work.

To motivate the idea behind representativeness, we consider an example in which an attack is strongly beneficial in terms of work, but not in terms of reputation. Consider the personalised PageRank algorithm as a reputation mechanism, which is trivially resistant to Sybil attacks in terms of reputation as the reputation scores it can assign are bounded by 1. However, the personalised PageRank algorithm is susceptible to strongly beneficial Sybil attacks in terms of work, as was shown by Liu et al. [5]. An example of such an attack would be the case where an attacker adds a single Sybil identity and creates two edges with very high edge weights connecting the attacker to its Sybil and vice versa.

5 ON THE IMPOSSIBILITY OF SYBIL-PROOFNESS

Using the more rigorous definitions of Sybil attack benefit, we now expand on the findings of Seuken and Parkes [11]. We briefly recap their results and point out an inaccuracy, which can be attributed to the absence of rigorous definitions for cost and profit. The authors make the following three assumptions (Definitions 5.1 - 5.3).

Definition 5.1 (Single-report Responsiveness, partly from [11])**.** Let $G_i = (V_i, E_i, w_i)$ be a subjective work graph with $k \in V_i$ such that there is no path in G_i connecting i and k. Now, let G'_i be the same subjective work graph as G_i, but with a directed path P of finite length connecting i and k with edge weights greater than some $c > 0$. We say a reputation mechanism R satisfies single-report responsiveness, if it holds $R_i(G'_i, k) > 0$.

Definition 5.2 (Independence of Disconnected Nodes, [11])**.** Given a subjective work graph $G_i = (V_i, E_i, w_i)$ with $l \in V_i$ such that $w_i(k, l) = w_i(l, k) = 0$ for all $k \in V_i$, let G'_i denote the subjective work graph of i, with $V'_i = V_i \setminus \{l\}$ and $w'_i(h, k) = w_i(h, k)$ for all $h, k \neq l$. A reputation mechanism R is said to satisfy *independence of disconnected nodes* if $R_i(G_i, k) = R_i(G'_i, k)$ for all $k \in V'_i$.

Definition 5.3 (Symmetry, [11])**.** Given a subjective work graph G_i, a reputation mechanism R is said to be *symmetric*, if for any graph isomorphism f with $f(i) = i$ it holds $\forall k \in V_i : R_i(G_i, k) = R_i(f(G_i), f(k))$.

These properties are not chosen arbitrarily, but are properties reasonable reputation mechanisms should satisfy. Definition 5.1 describes the necessity for a reputation mechanism to positively reward every contribution, while Definition 5.2 implies that passive agents should not influence the reputation. Definition 5.3 asserts that reputation scores should be independent of agents' identifiers. Using these assumptions, the authors prove the following theorem.

Theorem 5.1 (Impossibility of Sybil-Proofness, [11])**.** *For every reputation mechanism that satisfies independence of disconnected agents, symmetry and single-report responsiveness there exists a passive strongly beneficial Sybil attack, in terms of work.*

In their proof, the authors assume the following setting. Let G be a work graph with $i, j, k \in V$, such that j, who is malicious, has performed some work for the honest agent i, i.e., $w_i(i, j) > 0$. k is disconnected from the graph. If k now performs some work $c > 0$ for j then by single-report responsiveness it follows $R_i(G_i, k) > 0$. We assume that c is large enough that i will perform some work for k. Next, j creates a Sybil s, which by independence of disconnected nodes does not affect the reputation. Now one can apply a graph isomorphism to G_i swapping the labels k and s and as a result s will be able to consume some work. As no work was contributed in the attack, but some was consumed, the attack is strongly beneficial.

We disagree with this conclusion and argue that the cost of the attack was not zero, but larger than zero. We observe that the authors do not weigh the attack edge, $w_i(i, j) > 0$ into the cost of the attack. We argue that this edge constitutes a vital component of the attack and should be taken into account as part of the cost. With our definition of cost, their result does not hold, since the cost of the attack in their proof turns out to be larger than zero, making the attack not strongly but weakly beneficial at best.

Note that this result relies on a symmetry requirement for personalised reputation, as opposed to global reputation and therefore does not overlap with the results of Cheng and Friedman [2].

However, Theorem 5.1 can still be amended by introducing our requirement of parallel-report responsiveness.

Definition 5.4 (Parallel-report Responsiveness)**.** Let $G_i = (V_i, E_i, w_i)$ be an arbitrary subjective work graph with $j, k, l \in V_i$ such that there is no path in G_i connecting i and k, l and there exists a directed path P of finite length connecting i and j with edge weights larger than some $c > 0$. Now let G_i' be the graph G_i after k has performed some work for j, i.e., $w_i(j,k)' > 0$. Furthermore, let G_i'' be the graph G_i' after l has performed some work for j, i.e., $w_i(j,l)'' > 0$. We call a reputation mechanism R *parallel-report responsive* if it is single-report responsive and it holds $R_i(G_i', k) \leq R_i(G_i'', k)$.

Definition 5.4 suggests that if j adds multiple Sybil agents in parallel to one another, then the reputation of the Sybils will not be reduced by adding newer Sybils. We find that this definition is a common property of reputation mechanisms in the existing literature, such as BarterCast [6] and NetFlow [7].

Theorem 5.2. *Any reputation mechanism satisfying independence of disconnected nodes, symmetry, and parallel-report responsiveness has a strongly beneficial passive Sybil attack in terms of work.*

Proof. Let j launch a Sybil attack σ_S on G_i such that there exists a directed path P of arbitrary but finite length connecting i to j with edge weights larger than some $c > 0$. Let s_1 be disconnected. Due to independence of disconnected agents, this will not affect the scores of other agents. Now assume that s_1 performs some c' units of work for j resulting in the new subjective work graph G_i'. Then by single-report responsiveness, it will follow $R_i(G_i', s_1) > R_i(G_i, s_1)$.

In the next step, the attacker can create a second Sybil s_2 which by independence of disconnected nodes will again not affect the reputation scores of agents in V_i'. s_2 will also report having performed c' units of work for j, leading to a new subjective work graph G_i'', whereby s_2 will gain some reputation from the perspective of i due to single-report responsiveness. Due to the symmetry assumption and the fact that all edges in the Sybil region have the same weight, it holds $R_i(G_i'', s_1) = R_i(G_i'', s_2)$ and due to parallel-report responsiveness it also holds $R_i(G_i'', s_1) \geq R_i(G_i', s_1)$.

Consequently, the following holds for the profit of the Sybil attack in terms of reputation:

$$\omega^+_{\text{rep}}(\sigma_S) = 2 \cdot R_i(G_i'', s_1) \geq 2 \cdot R_i(G_i', s_1)$$

Inductively, it follows that increasing the number of Sybils to any arbitrary number m by the same paradigm, as indicated in Figure 3a, yields the profit

$$\omega^+_{\text{rep}}(\sigma_S) \geq (m+1) \cdot R_i(G', s_1).$$

If we assume that the reputation increase of the Sybils is large enough for them to receive work from i whenever they query i, as was the assumption of Seuken and Parkes [11] we have proved that

$$\lim_{|S| \to \infty} \omega^+_{\text{work}}(\sigma_S) = \infty.$$

Because there is only one attack edge $w_i(i,j) > 0$, we find that $\omega^-_{\text{work}}(\sigma_S)$ must be constant for all $m \in \mathbb{N}$. Therefore we conclude that the attack is strongly beneficial in terms of work. □

Theorem 5.2 has now delivered the impossibility result promised in [11] in accordance with our definitions of cost and profit. In the proof we made the assumption that the reputation increase of the Sybils would be large enough for them to be served some work by i, as was done by Seuken and Parkes [11]. However, we can also remove this assumption from the theorem above. In this case we will still be able to prove that, given our requirements, a reputation mechanism is susceptible to strongly beneficial Sybil attacks in terms of work, if we additionally assume that R_i is strongly representative. This is because even without the assumption of Sybils receiving work from i the attack above is strongly beneficial in terms of reputation. For the remainder of this paper, we will no longer make this assumption and instead prove the remaining results in terms of reputation scores and extrapolate them to benefit in terms of work using our definition of representativeness.

We call this type of attack *parallel attack*, illustrated in Figure 3a. Next, we introduce another analogous requirement with which we can prove the same assertion, called serial-report responsiveness.

Definition 5.5 (Serial-report responsiveness)**.** Let $G_i = (V_i, E_i, w_i)$ be a subjective work graph with $j, k, l \in V$ such that there exists no path in G_i connecting i and k, l and there exists a directed path P of finite length connecting i to j with edge weights larger than some $c > 0$. Let G_i' be the graph G_i after k performs some work for j, i.e., $w_i(j,k) > 0$. Let G_i'' be the same as G_i' after l performs some work for k with $w_i(k,l) \geq w_i(j,k) > 0$. We say that the reputation mechanism R is serial-report responsive if it is single-report responsive and the following two conditions are satisfied

$$R_i(G_i'', l) \geq R_i(G_i', k) \quad \& \quad R_i(G_i'', k) \geq R_i(G_i', k).$$

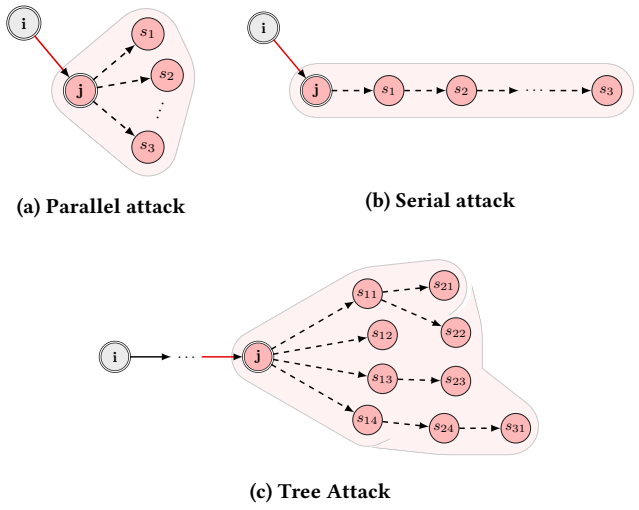
Definition 5.5 implies that if an attacker creates multiple Sybils in succession that all perform work for the previous Sybil, i.e., $w_i(s_n, s_{n-1}) > 0$, then the reputation of the Sybils is not influenced by adding newer Sybils and the newer Sybils will be assigned reputation values that are at least as large as the reputation of the older Sybils. We call this type of attack *serial attack*, shown in Figure 3b.

This property is also common among reputation mechanisms in the existing literature, such as in the BarterCast algorithm [6] and Netflow [7]. The ensuing Theorem 5.3 can be understood as a method of achieving the same result as Theorem 5.2 equivalent to parallel-report responsiveness and can be seen as another way of expanding on the incomplete theorem 1 from [11].

Theorem 5.3. *For every reputation mechanism R that satisfies independence of disconnected nodes and serial-report responsiveness, there exists a passive strongly beneficial Sybil attack in terms of reputation.*

The proof to this theorem follows analogously to the proof of Theorem 5.2 and we omit it due to spatial restrictions.

Trivially, there are many other requirements one could impose on reputation mechanisms to achieve the results above. The reason we chose the definitions of parallel- and serial-report responsiveness is that firstly, these are widespread properties of reputation mechanisms in the existing literature, and secondly, the two types of attacks in Figures 3a and 3b are the two elementary building blocks of any arbitrary Sybil attack. In Section 6 we argue that the profit of any arbitrary Sybil attack can be bounded by the profit of a combination of parallel and serial attacks, known as tree attack, visualised in Figure 3c.



(a) Parallel attack (b) Serial attack

(c) Tree Attack

Figure 3: Types of Sybil Attack

6 ON THE SYBIL-PROOFNESS OF REPUTATION MECHANISMS

In this section, we propose requirements for reputation mechanisms to be resistant to strongly beneficial Sybil attacks in terms of reputation. We begin by inverting the definitions of parallel- and serial-report responsiveness (Definitions 5.4 and 5.5) to bound the profit of parallel and serial attacks. We introduce a further requirement with which we show that the profit of an arbitrary Sybil attack can be bounded by the profit of a combination of the two elementary Sybil attacks, multiplied by a constant. Lastly, we ensure a positive cost of attacks, thereby achieving overall resistance to Sybil attacks.

Definition 6.1 (Parallel-report bound)**.** Let $G_i = (V_i, E_i, w_i)$ be a subjective work graph and let j be an attacker such that there exists a directed path P of finite length connecting i to j, with weights larger than some $c > 0$. Now let j launch a parallel Sybil attack σ_S with Sybil region $S = \{j, s_1, \ldots, s_m\}$. Without loss of generality we assume that it holds for the edges $w_i(j, s_l) = c_l \leq c_{l-1}$ for all $l \leq m$, i.e., we assume non-increasing edge weights, leading to the subjective work graph $G_i^{(m)}$. A reputation mechanism R is said to satisfy the **parallel-report bound** if it holds $R_i(G_i^{(m)}, s_l) \geq 0$ for all $l \leq m$ and for any $m \in \mathbb{N}$ we have

$$\sum_{l=1}^{m} R_i(G_i^{(m)}, s_l) \leq R_i(G_i^{(1)}, s_1).$$

This is not an arbitrary assumption, but one that can be found in most random-walk based reputation mechanisms, such as the PageRank algorithm [14] or the Hitting Time reputation mechanism [5]. It is an important property that prevents strongly beneficial Sybil attacks, as we will show in Lemma 6.1. However, it comes with the inherent trade-off that it assigns some honest agents smaller reputation scores than they would be entitled to and therefore limits a reputation mechanism's ability to induce cooperation.

Next, we introduce an analogous requirement for reputation mechanisms to be resistant to serial attacks.

Definition 6.2 (Serial-report bound)**.** Given the same conditions as in Definition 6.1, let j perpetrate a serial Sybil attack σ_S with Sybil identities $\{j, s_1, s_2\}$. A reputation mechanism is said to satisfy the **serial-report bound** if it holds for any two edge weights $w_i(j, s_1) = c_1, w_i(s_1, s_2) = c_2$

$$R_i(G^{(2)}, s_2) \leq R_i(G^{(1)}, s_1).$$

Definition 6.3 (Convergence of serial reports)**.** Given the same conditions as in Definition 6.2, a reputation mechanism R is said to satisfy **convergence of serial reports** if it holds for some arbitrary sequence $(c_l)_{l \in \mathbb{N}} \subset \mathbb{R}_{\geq 0}$ with $w_i(s_{l-1}, s_l) = c_l$, $R_i(G_i^{(n)}, s_l) \geq 0$ for all $l \leq n$ with a convergent sum

$$\lim_{m \to \infty} \sum_{l=1}^{m} R_i(G_i^{(m)}, s_l) < \infty.$$

Just like parallel-report bound, this is not an arbitrary definition tailored towards a certain result, but prominent property found in many random-walk based reputation mechanisms. It is important for the same reasons as Definition 6.1 and entails the same trade-off.

Definition 6.4 (Path-Responsiveness)**.** Given a subjective work graph $G_i = (V_i, E_i, w_i)$, we say that a reputation mechanism R satisfies path-responsiveness if $R_i(G_i, k) > 0$ implies that there exists a directed path P of finite length connecting i to k with non-zero edge weights.

Path-responsiveness prevents Sybil attackers from obtaining reputation scores ≥ 0 without performing at least some honest work through attack edges. It is a property satisfied by practically every reputation mechanism in the existing literature and simply enforces that any reputation score must be earned by performing honest work for the network.

Lemma 6.1. *Let G_i be the subjective work graph of honest agent i with attacker j launching a Sybil attack σ_S. Given a path-responsive reputation mechanism R_i, σ_S cannot be strongly beneficial in terms of reputation scores, if one of the following conditions holds:*

1. *If σ_S is a* parallel *attack and R_i satisfies the parallel-report bound.*
2. *If σ_S is a* serial *attack and R_i satisfies convergence of serial reports.*

We claim that the profit of any passive Sybil attack on any arbitrary graph structure can be bounded from above by attacks that are given by the combination of parallel and serial attacks. We refer to the combination of these two as *tree attacks*.

Definition 6.5 (Tree Sybil Attack)**.** Given an arbitrary objective work graph $G = (V, E, w)$ with malicious agent j. A passive Sybil attack $\sigma_S = (S, E_S, w_S)$ of arbitrary size $|S| = \sum_{i=1}^{M} m_i$ with $S = \left\{j, s_{11}, s_{12}, \ldots, s_{1m_1}, s_{21} \ldots, s_{2m_2} \ldots, s_{M1}, \ldots, s_{Mm_M}\right\}$ is called a tree Sybil attack if

$$\forall (j, s) \in E_S \,:\, s \in \left\{s_{11}, \ldots, s_{1m_1}\right\}$$
$$\forall 1 < l \leq M \, \forall i \leq m_l \exists! k \leq m_{l-1} \,:\, (s_{l-1k}, s_{li}) \in E_S.$$

An example of a tree Sybil attack is illustrated in Figure 3c.

Proposition 6.1. *Let G_i be a subjective work graph with attacker j launching a tree Sybil attack σ_S. If R_i satisfies convergence of serial reports, parallel-report bound and path-responsiveness, then σ_S cannot be strongly beneficial in terms of reputation.*

Proof. We begin by only examining the first *layer* of the tree, given by the Sybils $\{s_{11}, \ldots, s_{1m_1}\}$. The given tree confined to this layer is a simple parallel Sybil attack and we know by the parallel-report bound that it must hold

$$\sum_{k=1}^{m_1} R_i(G'_i, s_{1k}) \leq R_i(G_i, s_{11}).$$

The second layer of the tree attack can be interpreted as a number of Sybil attacks perpetrated by m_1 attackers. Again, we can apply the parallel-report bound and find that the profit of the second layer of each branch will be bounded by the profit of a serial attack with two Sybils . The serial-report bound ensures that the profit of the second layer will be bounded by the profit of the first layer, i.e.,

$$\sum_{k=1}^{m_2} R_i(G'_i, s_{2k}) \leq \sum_{k=1}^{m_1} R_i(G'_i, s_{1k}) \leq R_i(G_i, s_{11}).$$

We can continue this reasoning inductively and find that the profit of a tree Sybil attack with infinite layers, each containing finitely many Sybils, is bounded by the profit of an infinite serial Sybil attack. By convergence of serial reports, this profit will be finite. The only other way an attacker might attempt to obtain infinite reputation from a tree Sybil attack is by scaling one or more layers of the tree. Due to parallel-report responsiveness the profit of this attack is still finite and the attack cannot be strongly beneficial. □

If above weak representativeness is satisfied, the result holds in terms of work as well. Next, we introduce one additional property that bounds the profit of any passive Sybil attack by the profit of a tree attack multiplied by some constant. We call this property *multiple-path response bound*. We introduce the following operation to perform on a subjective work graph.

Let G_i be a subjective work graph with $k \in V_i$ such that there exist N directed paths $(P_n)_{n \leq N}$ connecting k to i. Now, define G'_i as an altered version of the subjective work graph of i, whereby the agent k is *split* into several agents $k_1, \ldots, k_N$, where every k_l ($l \leq N$) is connected to i by exactly one path. G'_i is created by splitting k into as many nodes as there are paths connecting it to i. We begin with k_1 and remove all agents and edges that are part of any of the paths $P_2, \ldots, P_N$ while keeping all which are part of P_1. We now relabel k (as the end-point of P_1), k_1. Next, we add path P_2 to the graph. Any agent j (or edge e) in P_2 that is also part of P_1, is now duplicated into j_1 and j_2 such that $j_1 \in P_1$ and $j_2 \in P_2$, i.e., ($e_1 \in P_1$ and $e_2 \in P_2$). We continue this for all paths $P_1, \ldots, P_N$ and obtain G'_i.

Definition 6.6 (Multiple-Path Response Bound)**.** Let G_i be a subjective work graph with $k \in V_i$ such that there exist N paths $(P_n)_{n \leq N}$ connecting i and k and let G'_i be the subjective work graph obtained by performing the operation above on G_i. We say that the reputation mechanism R satisfies the multiple-path response bound if it holds

$$R_i(G_i, k) \leq \sum_{n=1}^{N} R_i(G'_i, k_n).$$

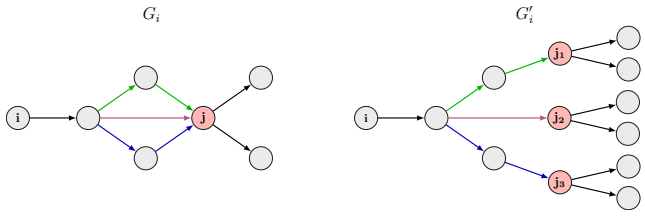


Figure 4: Example of multiple-path response bound applied to j from the view of i. j is connected to i via three paths and therefore split into three agents j_1, j_2, j_3.

Figure 4 and Figure 5 show examples of the multiple-path response bound applied to j in two different graph topologies. In Figure 4 there are three paths connecting i to j and applying the multiple-path response bound to j yields three agents, each connected to i via one path. In Figure 5 the graph contains a cycle. We interpret a cycle as infinite paths connecting i and j. Applying the multiple-path response bound graph transformation produces an infinite sequence of agents $(j_n)_{n \in \mathbb{N}}$ in G'_i where each j_n is connected to i via one path. This demonstrates that the given graph transformation can be applied to more sophisticated topologies as well.

As in our earlier restrictions on reputation mechanisms we claim that the multiple-path response bound property is not an arbitrary invention by us, but is in fact satisfied by all reputation mechanisms defined in [15] and plenty of the existing reputation mechanisms such as PageRank, Maxflow and Netflow [7]. Instead, we elaborate on the intuition behind this definition.

For a reputation mechanism to determine the cooperativeness of an agent, it needs to evaluate this agent's indirect contributions and consumption to/from i. i evaluates j by the incoming edges from i, whereby each path connecting j to i can be considered an indirect contribution and therefore, should influence the reputation score of j in i's subjective work graph. However, it is crucial for Sybil resistance that the effect of an additional path in the network should not exceed the effect that this additional path would have on $R_i(G_i, j)$ if it were the only path, as we do not want reputation to be gained disproportionately to the amount of work performed.

Next, we introduce transitive trust as a requirement to finally achieve Sybil-proofness.

Definition 6.7 (Transitive Trust)**.** Let G_i be a subjective work graph containing a directed path $P = (i, j_1, \ldots, j_n, j)$ of arbitrary length with strictly positive edge weights. We say R satisfies **transitive trust** if it holds

$$R_i(G_i, j_1), R_{j_1}(G_{j_1}, j_2), \ldots, R_{j_n}(G_{j_n}, j) > 0 \Rightarrow R_i(G_i, j).$$

We say that R satisfies **bounded transitive trust** if it also holds

$$R_i(G_i, j) \leq \min\left\{R_i(G_i, j_1), R_{j_1}(G_{j_1}, j_2), \ldots, R_{j_n}(G_{j_n}, j)\right\}.$$

If there are several (N) paths $(P_d)_{d \leq N}$ of lengths n_d given by $(i, j_1^d, \ldots, j_{n_d}^d, j)$ connecting i and j, then $R_i(G_i, j)$ must be bounded by the sum of the minimums given above (for each path).

$$R_i(G_i, j) \leq \sum_{d=1}^{N} \min\{R_{j_l^d}(G_{j_l^d}, j_{l+1}^d) \mid j_l^d \in P_d\}.$$

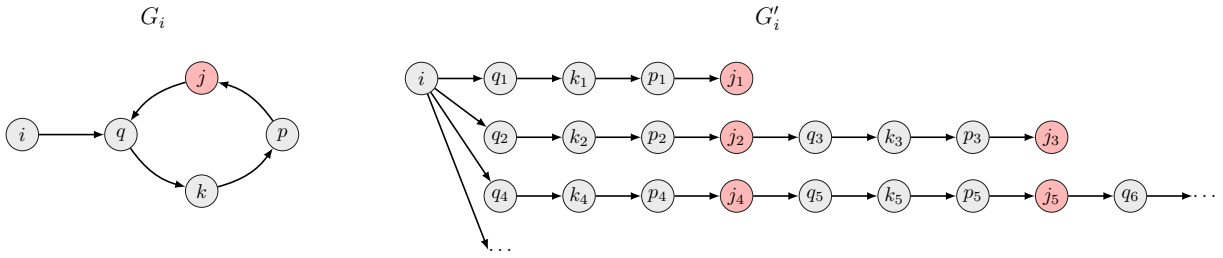


Figure 5: Example of multiple-path response bound applied to a graph with cycle. A loop implies infinite paths and therefore j is split into an infinite sequence $(j_n)_{n\in\mathbb{N}}$.

Definition 6.7 states that if i assigns j a reputation score greater than zero, and j assigns another agent k some reputation score greater than zero as well, then i assigns k some reputation score greater than zero as well. Bounded transitive trust implies that the reputation score k has with i must be bounded from above by the minimum of the reputation score i assigns j and the score j assigns k. It is a common property of reputation mechanisms, such as BarterCast and Netflow. Using these requirements we now prove Sybil-resistance of reputation mechanisms, as promised.

Lemma 6.2. *Let R_i satisfy the multiple-path response bound and bounded transitive trust. If j launches a passive Sybil attack σ_S. Then the profit $\omega^+_{rep}(\sigma_S)$ is bounded by the profit $\omega^+_{rep}(\sigma_{\tilde{S}})$ of a passive tree Sybil attack $\sigma_{\tilde{S}}$ multiplied by a constant $c < \infty$.*

Proof. First, assume there exists one directed path P connecting i to j. We can apply the multiple-path response bound to the Sybil region S yielding a new Sybil region $\tilde{S}$ in which every Sybil is connected to j via a single path. $\sigma_{\tilde{S}}$ is therefore a tree Sybil attack and the profit of the attack σ_S is bounded by the profit of $\sigma_{\tilde{S}}$. If there are finitely many directed paths $(P_n)_{n\leq N}$ connecting i and j then we can apply the multiple-path response bound to j, and obtain a subjective work graph G'_i with $j_1, \ldots, j_N$, each connected to i via a single path and committing the same Sybil attack. We obtain N equivalent Sybil attacks $\sigma_{S_1}, \ldots, \sigma_{S_m}$ and can apply the same procedure as we did in the case of a single path connecting i and j yielding N tree Sybil attacks $\sigma_{\tilde{S}_1}, \ldots, \sigma_{\tilde{S}_N}$. We can then infer the inequality $\omega^+_{rep}(\sigma_S) \leq N \cdot \omega^+_{rep}(\sigma_{\tilde{S}})$, where $\omega^+_{rep}(\sigma_{\tilde{S}})$ is the largest profit of the N tree Sybil attacks. Lastly, if there are infinite paths connecting i and j then the graph must contain a cycle and we can infer with the transitive trust property that $\sum_{n=1}^{\infty} R_i(G''_i, j_n) \leq \sum_{k\in N'(i)} R_i(G'_i, k)$, where N'_i is the neighbourhood of i in V'_i. Hence, we conclude analogously to the case of finite paths $\omega^+_{rep}(\sigma_S) \leq \omega^+_{rep}(\sigma_{\tilde{S}}) \cdot \sum_{k\in N'(i)} R_i(G'_i, k)$. □

Combining the results from Proposition 6.1 and Lemma 6.2, we argue that the profit of any arbitrary passive Sybil attack is finite.

Theorem 6.1. *Any reputation mechanism R satisfying path-responsiveness, multiple-path response bound, convergence of serial reports, the parallel-report bound, as well as bounded transitive trust is resistant to strongly beneficial passive Sybil attacks in terms of reputation.*

The proof follows directly from the proofs to Proposition 6.1 and Lemma 6.2. Using the properties of multiple-path response bound and bounded transitive trust, we obtain the following corollary.

Corollary 6.1. *Any reputation mechanism R satisfying path-responsiveness, multiple-path response bound, convergence of serial reports, the parallel-report bound, as well as bounded transitive trust is resistant to strongly beneficial active Sybil attacks in terms of reputation.*

Proof. Let σ_S be an active Sybil attack, then for any $c > 0$ we know there must be a bounded number of attack edges with edge weights larger than c. Therefore, we can apply the multiple-path response bound to each Sybil that is connected to an attack edge and obtain a finite number of passive Sybil attacks. The rest follows analogously to Theorem 6.1. □

If, in addition to the requirements stated above, R is weakly representative, then it is also resistant to strongly beneficial Sybil attacks in terms of work, as discussed in Section 4.

7 CONCLUSION

In this paper we have studied the Sybil-proofness of reputation mechanisms in multi-agent systems. We introduced rigorous metrics for the benefit of Sybil attacks, determined by the ratio of their cost and profit. While the goal was to bound the effect of Sybil attacks in terms of the work contributed and consumed by the attacker, these values were impractical to compute. We therefore introduced a pair of proxies, given in terms of the reputation obtained through the attack and through honest work. We introduced a requirement known as representativeness that ensures an equivalence between these two ratios. Using these metrics we revisited the impossibility result of Seuken and Parkes [11], pointing out an error which we attribute to ambiguity in their definitions of the attack benefit. We expanded on this result with two requirements we called parallel- and serial-report responsiveness and inverted the intuition behind these two requirements to obtain Sybil-resistance to parallel and serial attacks. We extrapolated our results to a combination of these two, known as tree attacks. Introducing a further requirement known as multiple-path response bound we achieve resistance to arbitrary attacks. Our bounds may seem loose, but a finite benefit ensures an attacker's contributions remain proportionate to its consumption, which is sufficient to protect the longevity of any multi-agent work system. In future work one may consider bounding the benefit of any attack by a fixed and finite value $c > 0$.

REFERENCES

[1] Ashwin R Bharambe, Cormac Herley, and Venkata N Padmanabhan. 2006. Analyzing and improving a bittorrent networks performance mechanisms. In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*. IEEE, IEEE, 1–12.

[2] Alice Cheng and Eric Friedman. 2005. Sybilproof reputation mechanisms. In *Proceedings of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems*. 128–132.

[3] Alice Cheng and Eric Friedman. 2006. Manipulability of PageRank under sybil strategies. In *First Workshop on the Economics of Networked Systems*. ACM.

[4] Sepandar D Kamvar, Mario T Schlosser, and Hector Garcia-Molina. 2003. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*. 640–651.

[5] Brandon K Liu, David C Parkes, and Sven Seuken. 2016. Personalized hitting time for informative trust mechanisms despite sybils. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1124–1132.

[6] Michel Meulpolder, Johan A Pouwelse, Dick HJ Epema, and Henk J Sips. 2009. Bartercast: A practical approach to prevent lazy freeriding in p2p networks. In *2009 IEEE International Symposium on Parallel & Distributed Processing*. IEEE, 1–8.

[7] Pim Otte. 2016. *Sybil-resistant trust mechanisms in distributed systems*. Master's thesis. TU Delft.

[8] Pim Otte, Martijn de Vos, and Johan Pouwelse. 2020. TrustChain: A Sybil-resistant scalable blockchain. *Future Generation Computer Systems* 107 (2020), 770–780.

[9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The pagerank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[10] Sven Seuken, Michel Meulpolder, DC Parkes, JA Pouwelse, J Tang, and DHJ Epema. 2014. Work accounting mechanisms: Theory and practice. In *Working Paper. Department of Informatics*. University of Zurich.

[11] Sven Seuken and David C Parkes. 2011. On the Sybil-proofness of accounting mechanisms. In *Workshop on the Economics of Networks, Systems, and Computation (NetEcon'11)*.

[12] Sven Seuken and David C Parkes. 2014. Sybil-proof accounting mechanisms with transitive trust. *Proceedings of the International Foundation for Autonomous Agents and Multiagent Systems* (2014).

[13] Sven Seuken, Jie Tang, and David C Parkes. 2010. Accounting mechanisms for distributed work systems. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[14] Alexander Stannat and Johan Pouwelse. 2019. A Random Walk based Trust Ranking in Distributed Systems. *arXiv preprint arXiv:1903.05900* (2019).

[15] Jie Tang, Sven Seuken, and David C Parkes. 2010. Hybrid transitive trust mechanisms. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems.

[16] Atsushi Yamamoto, Daisuke Asahara, Tomoko Itao, Satoshi Tanaka, and Tatsuya Suda. 2004. Distributed pagerank: a distributed reputation model for open peer-to-peer network. In *2004 International Symposium on Applications and the Internet Workshops. 2004 Workshops*. IEEE, 389–394.