Someone told me I can trust you: A graph-based reputation mechanism survey

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

Abstract

Keywords: Reputation mechanism, 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, governments may attempt to control news media to spread propaganda and manipulate their people. On the other hand, conspiracy theorists may try to spread their views on the world and society through the misuse of social media.

This paper aims to provide a survey of existing work in graph-based trust frameworks in a decentralized setting and suggest possible future work.

Purpose of this paper is to explore existing work in reputation mechanisms adopting trust graphs.

2 Background

Shaping trust in the online world, the main purpose of all reputation mechanisms, has always been a challenge. As the space of defense mechanisms gradually grows, so does the space of attacks. For example, as people are getting more aware of the risk of the internet and start to become sceptic towards (spam)mails, it is causing scammers to invent more intelligent and sophisticated scams [2]. Understanding trust is hard

- relentlessly evolving with the attackers for 27 years (ebay) - futureproofness
  https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3939644
  https://ieeexplore.ieee.org/abstract/document/7000170
  https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=ebay+fraud+analysis&btnG=
- The will to increase profits .... part about big tech corporations aiding the spread of misinformation
- TikTok has a very good recommender?
- red queen problem

'Big-tech' corporations are implicitly aiding the spread of misinformation with their current business models. For instance, YouTube's recommendation algorithm urges users to watch videos similar to those they have previously viewed. If a given person has viewed a video containing false information, YouTube's algorithm is likely to suggest other, potentially malicious, videos containing misinformation, as it increases the likelihood of that video being viewed and increasing revenue. More specifically, YouTube has attempted to battle this phenomenon during the COVID-19 pandemic by increasing the ranking of provaccine videos over antivaccine videos. It has however been found that antivaccine videos can still be recommended by YouTube when viewing provaccine videos [3].

Moreover, historical records show that the responsibility of creating trust can not 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 [4]. 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.

3 Definitions

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*" [5], where a trustor is the party which receives some service and the trustee is the party which is 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 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.

Not all entities will have experienced interactions with all other entities, therefore, for every entity, they can only assess the trustworthiness of a subset of all entities. Such relations

can be depicted in a directed graph, which we call a *trust graph*, which facilitates the necessary structural foundation. More specifically, we say 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 its trustworthiness, depending on the given reputation mechanism. Furthermore, entities can occur in multiple *trustsets*, but no entity can contain itself in its trust set: $\forall i \in N : i \notin N_i$ and $\bigcup_{i}^{n} N_i \cup \{j \in N \mid \forall k \in N : j \notin N_k\} = N$, where $n = |N|$. 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 $\forall(i,j,v),(k,l,w) \in E : \{(i = k \land j = l) \Leftrightarrow (i,j,v) = (k,l,w)\}$. Finally, all entities occur exactly once in a *trust graph*: $\forall i,j \in N : \{ID(i) = ID(j) \Leftrightarrow i = j\}$, where ID is a deterministic implementation-specific function capable of identifying individual entities. TODO SELECT THE USED SOURCES FOR TRUST GRAPHS

Assessment framework

Allowing us to ease comparison and break down existing solutions into their foundational components, we utilize a simple framework to detailedly describe essential aspects of such proposed solutions. These rigorous components consist of: *general strategy*, *mathematical foundation*, *resilience* and *Experimental performance*. Using such framework allows us to further identify a mechanisms fortes as well as it weaknesses.

General strategy – A mechanism's general strategy briefly elaborates on the intuitive ideas behind a reputation system. We convey to the reader the reasoning behind the discussed mechanisms and illustrate a perspective of why and how they work. The aim of the *general strategy* is to provide insightful readers with sufficient information to reproduce such reputation systems. When applicable, the *general strategy* will additionally clarify the mechanism's specific use case(s) and its bootstrapping protocol.

Mathematical foundation – All described mechanisms model a decentralized network using some variant of the aforementioned *trust graph*. The *mathematical foundation* describes the essential mathematical properties of the discussed mechanisms, as well as how the mechanisms can be mapped our *trust graph*. Some of the described mechanisms may introduce some additional ad hoc extensions or constraints to our generic model definition to suite a certain purpose.

Resilience – As a reputation mechanisms main purpose is to shape some form of status of benevolence within a network, it should be as resilient as possible. A well-known attack is the Sybil attack [6]. Defenses against this almost inexorable attack have been studied broadly TODO CITE. Another well-known attack type is the Eclipse attack [7] which attempts to isolate the target entity by targeting its incoming and outgoing edges. The *resilience* component discusses both the resilience against Sybil attacks, as well as other types of attacks we identified or identified by the authors of that particular mechanism.

TODO:

1. Further explain Sybil attacks
2. Problem definition

4 Existing mechanisms

Numerous approaches tackling the online trust problem, by introducing graph-based reputation mechanisms, have been proposed TODO CITATIONS. This section highlights a subset of methods proposed as a graph-based reputation mechanism for shaping trust. All discussed approaches are analyzed using the framework defined in section 3.

Souche

Souche is a vouch-based reputation mechanism developed partially by Microsoft[1] [8]. Its main goal is to quickly be able to distinguish between legitimate and illegitimate users in the context of online social communities. 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.

General strategy – Souche's main means for creating relationships between entities, i.e. users, is through implicit *vouching*. This implicit vouching process takes place by considering the

- "However, recent measurement studies indicate that two of the main assumptions on graph structures required by Sybil defenses, i.e., fast mixing social networks and the existence of a tight Sybil community, do not hold on real social graphs [22, 30]. The existence of compromised accounts further undermines those assumptions."
- For twitter social graph construction: if two people have mentioned eachother in tweets, there is an edge, so the graph is undirectional

Mathematical foundation –

- the graph is undirectional

Resilience –

VoteTrust

Introductory text

General strategy –

Mathematical foundation –

Resilience –

[1] https://microsoft.com/

[2] https://twitter.com/

PageRank

In the early ages of the internet, Google was among the first to adopt a graph-based reputation mechanism. Larry Page, Google's co-founder, introduced PageRank [9]: an algorithm used to rank pages based on relevance. While PageRank might no longer be Google's only reputation mechanism, it is the basis of many other reputation mechanism TODO SELECT CORRECT SOURCES.

General strategy – 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. Once the network reaches a stationary state, i.e. the rank does not change anymore, the 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.

Mathematical foundation – 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 be a function of web page p, such that:

$$R(p) = c \sum_{v \in B_p} \frac{R(v)}{|N_v|}$$

where B_p is the set of states $\{b \in N \mid p \in N_b\}$ 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 may occur with no outgoing edges, causing such loops to accumulate rank over time. To tackle this issue, Page introduced a new function R' of web page p such that $R'(p) = R(p) + cS(p)$, where $||R'||_1 = 1$, i.e. R' has a Manhattan distance of 1, and S(p) is a vector of web page p 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.

Resilience – The original version of PageRank as described above is prone to Sybil attacks, as has been shown in many studies [10–13]. 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' [12]. The original PageRank algorithm does by itself not contain any defense mechanisms against Sybil attacks. Furthermore, eclipse attacks are not applicable to PageRank as PageRank assumes that all nodes in the trust graph are known, whereas the eclipse attack assumes that nodes do not necessarily know of the existence of all other nodes.

MeritRank

MeritRank is a novel graph-based reputation mechanism which main goal is to define bounds for Sybil attacks [14]. That is, MeritRank does not attempt to solve Sybil attacks, but merely illustrates a number of strategies towards tolerating them. Furthermore, rather than performing actual computation, 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.

General strategy – 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 [15], 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.

Mathematical foundation – The aforementioned constraints are a set of intuitive measures to bound the gain of an adversary. Relative feedback/reputation limits the amount of reputation a node can give to some other node 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 a node 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 [9]) for reputation determination for any given node. Furthermore, connectivity decay defines a constant $0 \leq \beta \leq 1$ and ratio t, such that if for some node i (transitively) connected to some node j through some node k for at least the ratio t of all possible paths, $(1 - \beta)$ serves as a punishment factor for decreasing the reputation of the node 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 nodes to keep performing work to receive reputation.

Resilience – 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" [14]. 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 such an attack may muster.

EigenTrust?

Introductory text
General strategy –
Mathematical foundation –
Resilience –

Personalized Hitting Time?

Introductory text
General strategy –
Mathematical foundation –
Resilience –

MaxFlow

Introductory text
General strategy –
Mathematical foundation –
Resilience –

todo: more

https://www.usenix.org/legacy/events/nsdi09/tech/full_papers/tran/tran.pdf https://dl.acm.org/doi/abs/10.1145/1080192.1080202

1. Provide a nice table where different mechanisms are presented and labeled. Maybe also define a taxonomy?

5 Discussion

6 Conclusion

1. Summary
2. Say that much work has been performed in trust mechanisms and future work may focus on deploying vouching-based mechanisms in a decentralized setting.

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

[3] L. Tang, K. Fujimoto, M. T. Amith, R. Cunningham, R. A. Costantini, F. York, G. Xiong, J. A. Boom, C. Tao, *et al.*, ""down the rabbit hole" of vaccine misinformation on youtube: Network exposure study," *Journal of Medical Internet Research*, vol. 23, no. 1, p. e23262, 2021.

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

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

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

[7] E. Heilman, A. Kendler, A. Zohar, and S. Goldberg, "Eclipse attacks on {Bitcoin's}{peer-to-peer} network," in *24th USENIX Security Symposium (USENIX Security 15)*, pp. 129–144, 2015.

[8] 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.

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

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

[11] 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.

[12] G. Danezis and S. Schiffner, "On network formation,(sybil attacks and reputation systems)," in *DIMACS Workshop on Information Security Economics*, pp. 18–19, 2006.

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

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

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