Fighting misinformation with online trust: a literature survey

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. With the risk of misinformation and its impact rapidly becoming more vivid, the need of a universal trust framework arises. Such a trust framework would help enabling individual users to determine what information is true and whom to trust. In this paper, we aim to summarise the state-of-the-art approaches to creating trust in an online world.

Keywords: trust, online trust, trust framework, trust in web3, misinformation

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.

Furthermore, large corporations are implicitly supporting 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 [2].

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 [3]. 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.

As the risk of misinformation and its impact are rapidly becoming more vivid, so is the need for a universal trust framework, capable of helping users determine what and whom to confer trust. As early as 2002, researchers have pondered on and proposed methods for creating online trust [4]. This work proposes to model trust into three main components, namely trust, reputation and reciprocity, using a probabilistic mechanism for propagating and inferring these components. Another classic proposed solution utilizes user feedback to determine the trustworthiness of online articles [5]. This methodology aggregates user feedback to shape overall assessments of online resources, but relies on the user to perform additional work. With the transition to Web3, new opportunities for creating trust arise. Some of the proposed novel methodologies include the usage and verification of public records of interactions [6], registration of performed work in peer-to-peer networks [7] and using a blockchain's immutability for creating a robust reputation mechanism [8]. All the aforementioned methodologies are more thoroughly discussed and compared in section 3.

The main contributions of this paper are:

- Providing an overview of existing trust mechanisms.
- Presenting potential drawbacks or vulnerabilities these mechanisms may encompass.
- TODO: Providing an overview of defense mechanisms against Sybil attacks?
- TODO: Connecting trust mechanisms with Sybil attack defenses?

In this paper, we have summarised and reviewed existing work in the area of creating trust in an online world. Section 2 will provide more background information on trust. Secondly, in section 3 we will summarise different approaches to creating trust and discuss the dangers of Sybil attacks. Finally, ... TODO

2 Background and definitions

2.1 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*" [9], 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. However, for the remainder of this paper, we adapt the aforementioned definition to the following: *Trust is a Trustor's level of confidence in regard to the ability,* ***willingness and benevolence*** *of a Trustee to provide expected result in an interaction between Trustor and Trustee.*

Furthermore, Gambetta notes that another vital aspect of trust is the ability to disappoint [10]. Unless a trustee is not constrained in such a way that disappointment is non-viable, trust becomes irrelevant in the decision-making process, "for the more limited people's freedom, the more restricted the field of actions in which we are required to guess ex ante the probability of their performing them" [10]. In the context of computing systems, this statement entails that a trustee should have the ability to perform malicious actions or to disadvantage the trustor in anyway in order for a trust framework to become relevant.

Online trust can be decomposed into two main components: *content trust* and *entity trust*. Content trust focuses on whether or not the content of a certain resource can be trusted, independent of who has provided the resource. Trustworthy media may publish untrustworthy content. Entity trust focuses on trust between entities which can perform work for each other. However, these two types of trust are not necessarily independent, as an author's entity trust may (partially) be transferred into content trust for their provided resources depending on the underlying reputation mechanism. One may argue that if content and entity trust are to be considered independent, content trust can only be determined through (implicit) user feedback or advanced machine learning models.

While online trust enjoys much attention in the academic world, so do reputation mechanisms on which some trust frameworks are built. TODO: explain reputation mechanisms

Additionally, some of the discussed solutions utilize a trust graph. As the different trust graph-utilizing solutions propose different definitions of this data structure, we generalise its definition and map this to all discussed solutions, thereby easing comparison and aiding in outlining a clear overview. A trust graph is defined as a Directed Graph, composed of TODO find source which supports our approach (we can slightly adapt to fit the needs of this paper)

2.2 Sybil attacks

The Sybil attack [11] is a well-known strategy for abusing large distributed networks. An adversary employing a Sybil attack will generate numerous counterfeit identities and present these as distinct identities to the network. Such counterfeit identities may help an attack reach a number of goals, such as increasing a users reputation/trust by misleading the reputation mechanism deployed within the network or affecting the outcome of a majority vote within a distributed system.

Todo explain when circumstances when sybil attacks are possible.

Todo: extent this section depending on depth of literature survey into sybil attacks

3 Creating trust

This section discusses a variety of proposed solutions for creating trust and defending against Sybil attacks; their fortes and drawbacks will be discussed and compared. Lastly, we provide an overview of all discussed methods for the reader as a reference.

3.1 Entity trust mechanisms

Recent work has presented ConTrib as a mechanism for maintaining fairness among different entities in a distributed network by accounting work [7]. ConTrib assumes that all nodes create a digital signature for the communicated payload using their private keys. The public keys, which also act as unique identifiers, are communicated along with the payloads and signatures to ease verification by third-parties. In an effort to increase fraud resilience, ConTrib links records by including the incrementing record sequence number, the hash of the preceding record and pointers to pseudo-randomly determined prior records within the same personal database of records; the latter is used to speed up the verification process. Every message has to be answered by a confirmation message, containing the same information as well as the hash of the current message. Besides between both parties of the interaction, the messages are also communicated to f peers within the network. TODO discuss how this message protocol ensures fairness and allows for fraud detection. Also discuss the results of the 2 year Tribler experiment.

Creating trust through verification of interaction records.

Netflow? PageRank?

Trust metric

Maybe: Inferring reputation (partly) from the amount of money you have.

Maybe: If you misuse resources, you need to pay money, so you either collaborate or do nothing.

3.2 Content trust mechanisms

Trellis

Other stuff mentioned in introduction (todo).

3.3 Sybil defense mechanisms

4 Conclusion

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] 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.

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

[4] 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.

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

[6] J.-G. Harms, "Creating trust through verification of interaction records," 2018.

[7] M. de Vos and J. Pouwelse, "Contrib: Maintaining fairness in decentralized big tech alternatives by accounting work," *Computer Networks*, vol. 192, p. 108081, 2021.

[8] P. Peiris, C. Rajapakse, and B. Jayawardena, "Blockchain-based distributed reputation model for ensuring trust in mobile adhoc networks," in *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pp. 51–56, IEEE, 2020.

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

[10] D. Gambetta *et al.*, "Can we trust trust," *Trust: Making and breaking cooperative relations*, vol. 13, no. 1, pp. 213–237, 2000.

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