The Reputation Lag Attack

Doctoral Consortium

Sean Sirur
University of Oxford
United Kingdom
sean.sirur@gmail.com

ABSTRACT

Reputation systems and distributed networks are increasingly common. Examples include electronic marketplaces, online social networks, the Internet of Things [6] and ad-hoc networks (such as VANETs and MANETs) [2]. Such systems have great inherent complexity. As a result, they challenge typical methods for defining and verifying desired behaviour.

For example, actions such as leaving a smart home appliance open to network connections, connecting to a hotel WiFi network or interacting with a potentially pseudonymous identity (on an online marketplace or social network, for example) all entail an inherent component of uncertainty and, therefore, risk. As such, trust and reputation are increasingly crucial concepts for understanding modern socio-technical systems.

Broadly, we consider trust to occur whenever an agent interacts with another party without a guarantee of the other party's "good behaviour". A guarantee in this case depends on context. Information system guarantees are normally cryptographic whilst social guarantees can come in the form of face-to-face meetings; contracts; signatures; or "difficult to obtain/counterfeit information" such as bank cards, passports or driving licenses.

However, the users (or *trustors*) in systems with weakened or nonexistent guarantees may still attempt to protect themselves when considering an interaction. Commonly, this comes in the form of attempting to predict the future behaviour of potential interactive parties (or *trustees*).

For example, they may try to infer a trustee's future behaviour from their past behaviour. In doing so, they form an *opinion* of the trustee's *trustworthiness*. This distribution of opinions about a trustee constitutes their *reputation*.

However, a trustee's reputation may not reflect the reality of their past actions if there are time delays present in the system. This phenomenon is dubbed "reputation lag". Delays can come from a variety of sources and can manifest in different points in the trusting process.

If trustor's share trust-relevant information amongst themselves about trustees, the propagation of this information through the system may suffer delays. This kind of lag is referred to as *propagation lag*. This may be due to a variety of factors such as network connectivity, slow reporting and rating-update delays.

On the other hand, delays may be present in the evaluation of an outcome. A trustor's evaluation of their own interactions may also lag behind the reality of the trustee's behaviour. It may not be possible to (fully) evaluate a product's quality until some time has passed. Such cases of reputation lag are considered to be instances of "evaluation lag".

Again, there is a variety of potential causes for evaluation lag. This could include the inexperience of the trustor as a judge, the inherently "long-term" nature of a product's performance or, importantly, the trustee's deliberate attempts to conceal an interaction's negative consequences (e.g. Trojan viruses, or "catfishing").

A malicious trustee may attempt to intentionally exploit such lag to get away with self-serving bad behaviour that would not have been accepted if their reputation more accurately reflected their recent actions. Such a strategy is an example of a *reputation lag attack* [3, 4, 7, 10].

The classification of a behaviour as a reputation lag attack can be very broad, formally speaking. A behaviour or strategy constitutes a reputation lag attack simply if it results in the attacker acquiring more "gains" than they would have under perfect information.

There is evidence that reputation lag attacks could have a real negative impact on existing trust systems and proposed trust systems [5, 9]. Some preliminary investigations of the attack's impact in comparison to that of others have been performed. However, there has been no in-depth formal analysis of the reputation lag attack itself and the existing informal analyses introduce extraneous artefacts into their findings and conclusions.

In this thesis, we present a pair of fundamental but complementary models to capture both the phenomenon of reputation lag and its exploitation by a malicious attacker trustee.

The first model consists of a continuous time Markov chain (CTMC) that describes the sequences of events in the system [8]. An event can be an action by the attacker or a communication between users (who are nodes of a network) in which they share the outcomes of their interactions with the attacker.

This formal model is used to capture the core properties of the attack: firstly, the reputation of an actor failing to reflect their behaviour due to lag and, secondly, a malicious actor exploiting this for their personal gain. The formal model provides insight into the attacks, even without data.

In the first content chapter, we investigate this formal model. Three primary insights were gained from the model.

There exists a superior ordering to the attacker's actions: the attacker behaves positively; waits for this reputation to spread to as many users as possible; then begins behaving as negatively as possible before the users reject them. However, it was crucial that the users' judgement of the attacker's prior behaviour did not have a "decay factor" (i.e. for ordering to work, the users most not distinguish between more and less recent actions). Trust/reputation with decay factors may be somewhat more resistant against these attacks.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

Next, when dealing with an optimally dealing attacker, increasing user communication rates cannot be detrimental to users and may be detrimental to the attacker. Intuitively, this follows directly from the definition of the reputation lag attack which relies on poor user communication.

Finally, we showed that finding the sequence of attacker actions that performs the maximum number of cheats is an NP-hard problem. This is true even when the attacker has perfect foresight of the time of their action opportunities and of the users' communications.

For the second content chapter, a simulator of the above formalism was coded in Python. There were two main aims to using the simulator. The first aim was to try and illustrate the quantitative impact of the previous qualitative results. That is, we investigated how *much* rate, *deals* and ordering impact the attacker's success.

The second aim was to investigate more concrete aspects of the system: the impact of some concrete network structures between the users and the benefit to the attacker of being able to see each user's trust state before committing to a particular target for each action.

The attacker's rate was the dominant factor in their success. A sufficiently high rate was also beneficial if not necessary to the success of most other attack strategies.

We studied the attacker's ability to improve their reputation using *deals* and to wait for the users to spread existing information around the network. Randomly performing *deals* did not provide any unfair benefit. Randomly waiting was detrimental to the attacker, allowing users to rapidly spread knowledge about the attacker's *cheats* through the network.

However, we then demonstrated that by enforcing the "*deal* then wait then *cheat*" ordering to their actions, the attacker could outperform all previous strategies. Furthermore, if a sufficiently fast attacker waits for the mean duration of just a single *deal* to spread through the network, they begin to perform the maximal number of *cheats*. The difference is that in the former case, mostly negative reputation spreads in the wait periods, whereas in the latter, purely positive reputation spreads in the wait period.

Third, we investigated the efficacy of using network centrality measures to rank users on how likely they were to spread information through the system. The attacker would then focus their *cheats* on users with a low rank. We found that, while the ranking does improve the attackers success, it is a somewhat marginal improvement and that this is likely due to the lack of centrality measures that correctly capture this model's infection-style spreading [1].

We also showed that this strategy is reliant on the network having a suitably non-homogeneous structure and that, otherwise, the centrality measures fail to give distinct values for the users. However, this investigation was performed on "community-style" random networks (i.e. connected Watts-Strogatz and Barabási-Albert models) and these are both highly connected and relatively homogeneous with respect to the centralities used.

Finally, we tested an attacker with "*clairvoyance*" (i.e. the ability to access users' trust states before acting). This resulted in modest but clear improvements in attackers at medium rates as such attackers are the most likely to suffer due to repeated rejections.

In the third content chapter, we introduce a model that is primarily made up of a dynamical system in which the gains, behaviour and reputation of the attacker are interdependent functions in time.

The main difference of this model is that the behaviour and behavioural dependencies in the system (e.g. attacker quality as a function of attacker actions and reputation as a function of attacker quality) are defined explicitly. Behaviour which reflects that of the desired system to be modelled can be directly defined as part of the model.

The benefit of this is that there is no need to define or assume particular attack or trust mechanisms, preventing the model from becoming overly specific or restrictive. The downside is that the chosen functions must be robustly justified either through external experimentation/data-gathering or reliance on results found in related work.

In the final content chapter, we discuss and investigate some potential methods to mitigate the reputation lag attack. There are a few such potential solutions to the RLA.

A time decay of knowledge weighting can be introduced. For example, the users can discount the importance of older good actions to disrupt ordering.

Limiting the attacker action rate could prevent such attacks due to the apparent necessity of high attacker rates for effective RLAs.

Limiting the number of simultaneous interactions that the attacker can be involved in or the number of interactions that can be performed in a given period is a similar but distinct solution to rate limitation. This could, for example, prevent RLAs whilst allowing honest trustees to handle normal rapid behaviours like large influxes of requests (e.g. a seller during a clearance sale) whilst preventing that behaviour becoming endemic to their strategy as is necessary to the RLA. However, with clever timing of their waiting period, an ordered attacker could still perform an RLA depending on the system).

The presence of centralised mechanisms for reputation and interaction (e.g. Amazon, eBay) could lower the risk of propagation lag but this also comes with drawbacks as it introduces a single-point of failure. It also doesn't tackle evaluation lag.

The primary contributions of this thesis are as follows:

A high rate of attacker actions relative to the rest of the system is shown to be imperative to their success.

Under the right conditions (e.g. no time decay factor on reputation and the attacker's lifetime in the system is finite), the attacker benefits from an "ordering" strategy. In this strategy, they begin by behaving well, then they wait for their good reputation to spread around the system and then they begin behaving badly until they are rejected from the system.

Targetting users based on their position on the network is not of great benefit to an attacker and can in fact be detrimental at high rates, which is when they stand to gain the most.

An attacker benefits from having access to current user trust states (i.e. whether a user will reject them or not) but mostly at medium rates.

KEYWORDS

Reputation System; Social Simulation; Reputation Lag; Attacks

ACM Reference Format:
Sean Sirur. 2022. The Reputation Lag Attack: Doctoral Consortium. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

REFERENCES

[1] Stephen P Borgatti. 2005. Centrality and network flow. *Social networks* 27, 1 (2005), 55–71.

[2] Fogue et al. 2015. Securing Warning Message Dissemination in VANETs Using Cooperative Neighbor Position Verification. *IEEE Transactions on Vehicular Technology* 64, 6 (2015), 2538–2550. https://doi.org/10.1109/TVT.2014.2344633

[3] Audun Jøsang and Jennifer Golbeck. 2009. Challenges for robust trust and reputation systems. In *Proceedings of the 5th International Workshop on Security and Trust Management (SMT 2009), Saint Malo, France*, Vol. 5. Citeseer.

[4] Reid Kerr and Robin Cohen. 2006. Modeling Trust Using Transactional, Numerical Units. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services* (Markham, Ontario, Canada) *(PST '06)*. Association for Computing Machinery, New York, NY, USA, Article 21, 11 pages. https://doi.org/10.1145/1501434.1501460

[5] Reid Kerr and Robin Cohen. 2009. Smart cheaters do prosper: defeating trust and reputation systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 993–1000.

[6] Liu X., Abdelhakim, M., Krishnamurthy, P., Tipper, D. 2018. Identifying Malicious Nodes in Multihop IoT Networks Using Diversity and Unsupervised Learning. (2018), 1–6. https://doi.org/10.1109/ICC.2018.8422484

[7] Tim Muller, Yang Liu, Sjouke Mauw, and Jie Zhang. 2014. On robustness of trust systems. In *IFIP International Conference on Trust Management*. Springer, Springer Berlin Heidelberg, 44–60.

[8] Norris, J. 1997. *Continuous-time Markov chains I.* Cambridge University Press, 59–111. https://doi.org/10.1017/CBO9780511810633

[9] Carl Shapiro. 1982. Consumer information, product quality, and seller reputation. *The Bell Journal of Economics* (1982), 20–35.

[10] Sean Sirur and Tim Muller. 2019. The Reputation Lag Attack. In *Trust Management XIII*, Weizhi Meng, Piotr Cofta, Christian Damsgaard Jensen, and Tyrone Grandison (Eds.). Springer International Publishing, Cham, 39–56.