Identifying Vulnerabilities in Trust and Reputation Systems

Taha D. Gunes, Long Tran-Thanh, Timothy J. Norman

18th Aug 2019

Appeared in
Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)

1 Problem Addressed

To evaluate trust and reputation systems against known attacks, By presenting a method to automatically identify vulnerabilities in existing trust models. To provide reliable and objective means to assess how these systems are towards different kinds of attacks.

2 Previous Work

- Numerous kind of attacks and defence strategies have been explored [Hoffman et al., 2009], but considered relatively simple attack profiles.
- The BRS (Beta Reputation System) with filtering [Whitby et al., 2004], focuses on excluding attackers who provide unfair feedback by badmouthing or ballot-stuffing.
- The TRAVOS [Teacy et al., 2006] discounts outlying ratings in making trust assessments.
- The HABIT [Teacy et al., 2012] model uses a hierarchical Bayesian model to identify participants with various profiles of reliability, and factor into aggregated ratings.
- Bidgoly & Ladani [2016] considered injecting false evidence and whitewashing, which are modelled as primitive actions in a planning mechanism (POMDP).

3 Contributions

The contributions made here are three-fold.

- **First**: Model coordinated, strategic attacks with a specific objective as a derivative-free optimization problem.
- **Second**: Two search methods are proposed for efficiently identifying coordinated attacks in complex attack spaces through sampling-based optimization.
- **Third**: This method is used to analyze a selection of existing trust models, providing evidence for the kinds of complex attacks they are vulnerable to.

3.0.1 Basic Assumptions

- Prediction of the future behaviour of an agent (i.e. a trust assessment) at time t is, $\varepsilon = \{O^{0:t}_{c_i->p_i} | c_i \in C, p_i \in P\}$
- We investigate cases in which an attacker is limited by:
  - *Power*, the number of observations that it can add through the attack ($\rho = |\varepsilon^{'}|$)
  - *Control* over the witnesses ($W^{'} \subseteq W$).

3.0.2 Attack Space

- The space of possible attacks is χ,

$$|\chi| = \begin{cases} \rho + k.|\{O^{0:t}_{w_i->p_j} | w_i \in W^{'}, p_i \in P\}| - 1 \\ k.|\{O^{0:t}_{w_i->p_j} | w_i \in W^{'}, p_i \in P\}|. \end{cases}$$

- The space of attacks is defined in terms of:
  - The number of witnesses to be used, s
  - The distribution of the attack power, ρ across these selected witnesses, considering those they can report on:
    i. All restricted partitions of ρ into s ($D = RP_s(\rho)$) and their permutations without repetition: ${P_s}^D$
    ii. The distribution of these permutations to each witness-provider pair, such that the number of possible distributions is $(|P|.k)^s$
- The number of attacks in reduced space is,
  $|\chi| = \binom{|W^{'}|}{s} D.{P_s}^D.(|P|.k)^s$
- To solve attackers optimisation problem, 'Monte Carlo Sampling' or 'Hierarchical Sampling' based techniques are used.

4 Conclusion

A novel method for identifying vulnerabilities in trust and reputation systems is introduced and its practical value is demonstrated. Model when employed to search for effective strategies through derivative-free optimization methods, output a set of attack profiles and an estimate of the vulnerability of the TRS to an attack of that kind.