Exploring Trust-Based and Reputation-Aware Frameworks for Attack Detection in Recommender Systems: A 5-Year Survey

Dr. Priyanka Mishra, Prof. Shashi Kant Gupta
pdf.priyankamishra@lincoln.edu.my, shashigupta@lincoln.edu.my
Course: PDF in Computer Sc. & Emg
Lincoln University College, 47301 Petaling Jaya, Selangor Darul ,Ehsan, Malaysia

Abstract

Digital platforms use recommender systems to increase user engagement by personalizing content for users. Because of this, they are often abused by outside influences that will manipulate the system to favour themselves – as in the case of profile injection or shilling attacks. In recent years, there has been a shift to use of non-statistical means of detection, such as relational credibility, trust propagation, temporal consistency, and coordinated behavioural patterns as ways to identify suspicious users. This literature review provides a critical assessment of the progress that has been made in areas of trust-based/reputation-aware attack detection for recommender systems, summarizing key methods and identifying their strengths and weaknesses while offering some potential future directions. The review finds that hybrid models using trust, reputation, temporal signals and collusion analysis provide a greater degree of robustness than traditional rating based detectors, but also suggests that significant challenges exist with regard to cold-start users, explainability/fairness, real-world deployment and the threat of AI-generated stealth attacks. The paper concludes that research into the security of recommender systems should focus on dynamic trust modelling, graph-based detection and interpretable defence mechanisms.

Keywords: recommender systems, shilling attacks, trust-aware systems, reputation-aware frameworks, poisoning attacks, attack detection, collaborative filtering

1 Introduction

Recommender systems are one of the basic building blocks of all modern digital platforms, including e commerce, streaming, online education and social media. These systems provide a way to personalize the user experience by predicting preferences through previous interactions. However, despite the benefits of recommender systems, they are highly susceptible to malicious attack. One such attack is Profile Injection or Shilling or Poisoning Attack, in which an attacker creates fabricated ratings/profile to artificially promote or demote items (Zhou et al., 2018; Yang and Niu, 2021).

Attack detection was traditionally based on handcrafted statistical indicators (e.g., rating deviation, filler size, variance pattern). Although these types of attacks were generally detected using traditional detection techniques, it has become increasingly difficult due to the attacker's ability to mimic & quot; normal & quot; user behaviour. As a result, researchers are now mov-

ing toward trust and amp, reputation-based attack detection frameworks that assess not only how users rate an item, but how credibly and consistently they behave in the recommender system environment (Jiean et al., 2019; Xu et al., 2024).

This paper will provide a brief critical review of the literature on trust-based and reputation-based frameworks for attack detection in recommender systems over the past five years. It examines the emergence of these approaches, compares their contributions, and identifies the main methodological and practical gaps that remain.

2 Background: Attack Detection in Recommender Systems

Collaborative filtering is the basis for recommender system recommendations, which are based on user-item interaction history. Attackers can create fake profiles and assign ratings for targeted and filler items in order to manipulate these systems, as there are many opportunities for manipulation when using only user-generated ratings and behavior as input data (Zhou et al., 2015, 2018).

Traditional methods for detecting this type of attack have focused primarily on the use of statistical rating anomalies. Unfortunately, these approaches often suffer from a lack of ability to detect attacks when malicious users behave similarly to legitimate users. As a result of these shortcomings, researchers have begun to explore richer defense strategies that incorporate additional sources of information such as trust, reputation, temporal patterns, and relational consistency.

3 Trust-Based and Reputation-Aware Detection Frameworks

3.1 Trust as a Detection Signal

The concept of trust in a recommendation system generally refers to how trustworthy (or credible) one user is viewed by another user or group of users. Trust can be either explicit (i.e. social trust links) or inferred from the consistency of interactions, similarity, and the quality of historical interactions. Since malicious users often show themselves to be malicious based upon weak, inconsistent, or strategically query-manipulated trust relationships, trust in security-oriented recommendation systems is important (Liu et al., 2022; Trust in recommender systems, 2026).

An important study in this area was done by Jiean et al. (2019), who proposed a method of detecting suspicious users in a social recommendation system through combining time series analysis with trust features. Their proposed framework demonstrated that the use of trust-aware signals was more effective than the use of ratings-only for the identification of suspicious users, especially in environments where the recommendations were made through social connections.

3.2 Credibility Assessment through Recurring Reputation

Trust is usually dependent upon relationships while reputation can be seen as an indication of an individual's or an organization's long-term credibility or reliability, which gives insight into long-term credibility and consistency for a user. A reputation-aware detection model is attempting to identify patterns of consistent user behaviour across time and determine if those behaviours conform to a user's authenticity through the use of authentic recommendations.

Zhou et al. (2018) developed a detection framework that analysed the credibility of group users based on group users and rating time series. Their results concluded that, while users may behave maliciously, this behaviour is often identified through the identification of patterns of group activity rather than isolated individual behavioural anomalies. This may be why a reputation-aware detection model is applicable for discovering suspicious collective user activity.

3.3 Hybrid Trust-Reputation Models

In support of the growing trend toward hybrid models, more researchers continue to explore hybrid models that integrate trust, reputation, temporal analysis and contextual user behaviour. Hybrid models are therefore stronger than traditional approaches because they are looking for multiple complementary indicators to arrive at conclusions versus the once-seen value of one statistical measurement.

For example, Yang and Niu (2021) proposed a genre trust model wherein users are evaluated within the scope of the particular context for which a recommendation is provided; therefore, credibility is context dependent and not a global property of the user. Similarly, Xu et al. (2024) emphasised the importance of multidimensional variables of credibility characteristics of users as well as the analysis of collusive behaviours to detect group collusion attacks.

4 From Classical Shilling to Modern Poisoning Attacks

The manipulation of recommender systems has become increasingly covert and proficient, with attackers deploying machine learning techniques and generative programming to create valid-looking profiles to imitate legitimate user behaviour.

Lin et al. (2022) showed that even black-box recommender systems can be compromised by using generated, fake profiles, while Huang and Li (2023) demonstrated that even a single user can perform undetected, single-user invisible injection attacks that cause significant disruption to system recommendations. These results present new challenges to the old static trust and reputation models and provide further justification for adapting/adapting defence mechanisms.

Research from previous years has also shed light on a critical insight: not all users are equally impacted by attack incidents. Shrestha (Spezzano & Pera) maintains that users who have sparse history or niche preferences will be at a higher risk of manipulation and underlines the need for user-centered assessment of robustness.

5 Critical Evaluation and Future Directions

Over the last 5 years there has been substantial progress in the development of trust-based and reputation-based mechanisms to detect potential attacks, that better identifies relational inconsistencies, coordinated behaviour amongst groups, and long-term credibility patterns than earlier phases of statistical detection mechanisms. Despite all the recent progress research in this area, there still remains a number of significant constraints on the research.

One of the most significant limitations is the lack of a standardised means of conducting evaluations; different studies do not conduct evaluations using the same dataset, attack scenarios, or performance measures; therefore direct comparisons to other studies becomes very difficult.

Second, the freshly created frameworks are primarily validated through simulation and benchmark systems and rarely validated through real-world recommender systems, therefore there are issues regarding scalability and the practical application of these new frameworks.

Third, there are still a number of studies that have conducted a mathematical model of trust and reputation, but the explanation as to how the study classified the user as suspicious is not made clear. This negatively impacts transparency and diminishes the ability for the studies to have any practical relevance, and is detrimental to the reliability of the whole framework.

Finally, the issue of fairness still remains a great concern, especially for cold start users and for minority users who may be unfairly penalised by credibility based systems.

Future research should study the parameters of creating trust dynamically, developing new graph-based methods to help detect computer-generated deception attacks, applying improved explanatory data security mechanisms, and developing more effective defenses against computer-generated deception attacks produced by artificial intelligence (AI). Furthermore, there is a requirement for more realistic and privacy-preserving benchmark datasets to support repeatable results and practical relevance.

6 Conclusion

Trust-based and reputation-aware frameworks are emerging as an effective defence against attacks using recommender systems. There has been a clear transition from using simple rating-based methods for detecting outliers to more contextually aware and multi-faceted methods that highlight credibility.

Trust-based models will assist with detecting inconsistent relationships, and reputation-aware models will further assist with detecting outliers through long-term behaviour assessment. Hybrid approaches including trust, reputation, temporal signals, and collusion detection will provide the greatest promise.

Nonetheless, ongoing challenges exist within the environment such as fairness, cold start, and explainability; and with the existence of an increasingly sophisticated computer-generated deception attack landscape. Accordingly, as recommender systems continue to influence digital

decision-making, security among recommender systems must become increasingly accurate, while also being more credible, adaptable, and deployable in practice.

References

Huang, C. and Li, H. (2023) 'Single-User Injection for Invisible Shilling Attack against Recommender Systems', *arXiv*. Available at: https://arxiv.org/abs/2308.10467.

Jiean, W., Wen, J., Gao, M., Xiong, Q. and Koh, Y.S. (2019) 'Detecting shilling attacks in social recommender systems based on time series analysis and trust features', *Knowledge-Based Systems*, 178, pp. 25–47.

Lin, C., Chen, S., Zeng, M., Zhang, S., Gao, M. and Li, H. (2022) 'Shilling Black-box Recommender Systems by Learning to Generate Fake User Profiles', *arXiv*. Available at: https://arxiv.org/abs/2206.11433.

Liu, Z., et al. (2022) 'A survey for trust-aware recommender systems: A deep learning perspective', *Knowledge-Based Systems*, 249, p. 108954.

Shrestha, A., Spezzano, F. and Pera, M.S. (2018) 'Who is Really Affected by Fraudulent Reviews? An analysis of shilling attacks on recommender systems in real-world scenarios', *arXiv*. Available at: https://arxiv.org/abs/1808.07025.

'Trust in recommender systems: A survey' (2026) *Expert Systems with Applications*, 298, p. 129653.

Xu, Y., Zhang, P., Yu, H. and Zhang, F. (2024) 'Detecting Group Shilling Attacks in Recommender Systems Based on User Multi-dimensional Features and Collusive Behaviour Analysis', *The Computer Journal*, 67(2), pp. 604–616.

Yang, L. and Niu, X. (2021) 'A genre trust model for defending shilling attacks in recommender systems', *Complex & Intelligent Systems*, 9, pp. 2929–2942.

Zhou, W., Wen, J., Koh, Y.S., Xiong, Q., Gao, M., Dobbie, G. and Alam, S. (2015) 'Shilling Attacks Detection in Recommender Systems Based on Target Item Analysis', *PLOS ONE*, 10(7), e0130968.

Zhou, W., Wen, J., Qu, Q., Zeng, J. and Cheng, T. (2018) 'Shilling attack detection for recommender systems based on credibility of group users and rating time series', *PLOS ONE*, 13(5), e0196533.