Privacy Leakage and Quantitative Analysis in Social Network Big Data

Chengtian Zhu*

Engineering Department, Penn State University, State College, Pennsylvania 16801, United States

Abstract. With the development of social networks, multiple user data not only creates opportunities for data-driven applications, but also brings a serious privacy breach risk. Research now always focuses on qualitative description and lacks a systematic quantitative assessment framework. This article uses some open data sets like Twitter and Facebook, which model three typical attack scenarios--attribute reference, linkage predictions, and model reversals--and rates the anonymous, differential privacy, mixed defense, and federal study. The result of experiments says that, although in an anonymous situation, privacy hazards are still notable. Attackers can refer to sensitive information from the structure and behavior features. Although anonymity decreases the attack mission rate, distinct utility losses always occur. Mixed defense realizes a better balance. Federal studies have the ability to resist model reversal and face the challenge of expansion. Further analysis shows that cross-platform datasets and the activity of users magnify the privacy risk. In conclusion, this research not only creates new strategies in qualitative privacy leaks but also has a balance in data effectiveness and privacy protection between social platforms and regulators.

1 Introduction

Now, user activities across various platforms are continuously monitored and recorded, including content posted on social media, social relationships, geographic locations, daily activities, and personal interests. The accumulation of this data has created vast repositories of social big data, which hold significant value in areas such as public security, social governance, and scientific research. However, while presenting great opportunities, social network big data also poses serious risks of privacy leakage [1].

Previous studies have demonstrated that even anonymized social network data can be vulnerable to re-identification attacks through structural features or cross-platform data linkage [2]. Additionally, sensitive personal information—such as gender, age, contact numbers, and personal identifiers—can be indirectly inferred through attribute inference [3]. Furthermore, analyses based on user behavioral patterns may lead to more severe privacy threats, including behavioral modeling and profiling. These risks not only jeopardize individual privacy and security but also pose significant challenges to social trust and data sharing.

* Corresponding author: cqz5268@psu.edu

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

A major limitation of existing research is its reliance on qualitative descriptions and the lack of systematic frameworks for quantitative analysis. The scope and extent of privacy leakage are difficult to measure accurately, which restricts meaningful comparisons between different attack methods and impedes rigorous evaluation of defense mechanisms. Although established protection methods—such as k-anonymity, l-diversity, and differential privacy—can mitigate risks to some degree, they often do so at the expense of data utility [3]. Furthermore, research on cross-platform privacy leakage and the temporal evolution of privacy risks remains insufficient [4].

To address these challenges, this study makes several contributions. First, it proposes a unified framework for quantifying privacy leakage, thereby standardizing related concepts and enabling systematic comparisons. Second, it introduces metrics such as mutual information and re-identification probability to measure privacy risks in social networks. Third, it conducts empirical experiments on multiple publicly available datasets (e.g., Twitter and Kaggle), evaluating the effects of attacks, including attribute inference, link prediction, and model inversion. Additionally, the study designs and evaluates a hybrid defense approach that integrates differential privacy with graph structure preservation, aiming to enhance privacy protection while maintaining high data utility. Finally, the analysis is extended to dynamic networks and cross-platform data fusion scenarios, exploring how privacy risks evolve over time and across aggregated datasets.

Through these contributions, this paper not only fills a gap in the existing literature on the quantitative analysis of privacy leakage in social networks but also provides practical insights for social platforms and regulators. The findings offer both theoretical foundations and practical implications for ensuring the secure and sustainable use of social big data in the future.

2 Related Work

Extensive research has shown that social media and social network data are particularly vulnerable to various types of privacy attacks [5]. Narayanan and Shmatikov demonstrated that even anonymized social network data remains insecure, as de-anonymization can be achieved by exploiting structural similarities with auxiliary datasets, leading to large-scale re-identification attacks. Kosinski et al. found that users' sensitive information—such as gender, age, and personal preferences—can be inferred from their behavioral traces. Sala et al. proposed a differentially private graph model for data sharing but noted that it still has limitations in preventing re-identification [6]. More recently, model inversion attacks against graph neural networks (e.g., GraphMI) have revealed that even trained models may inadvertently leak structural and attribute information.

To systematically assess privacy risks, various measurement approaches have been proposed. Information-theoretic methods, such as entropy and mutual information, are widely used to quantify reductions in the uncertainty of sensitive attributes [7]. Graph-based anonymization models, including k-anonymity and l-diversity, provide formal privacy guarantees at both the structural and attribute-distribution levels. The differential privacy framework offers stronger mathematical guarantees by introducing noise into query results or graph structures. Recently, researchers have proposed unified quantitative frameworks that aim to evaluate privacy risks across different attack scenarios and facilitate cross-comparisons. However, these studies remain fragmented and lack consistency across various contexts [8].

Existing privacy protection mechanisms in social networks can generally be categorized into three types. The first category includes anonymization and perturbation techniques, such as node or edge deletion, attribute generalization, and structural perturbation [9]. The second category comprises differential privacy methods, which are widely applied in network data

publishing and recommendation systems but often compromise data utility. The third category involves decentralized approaches, such as federated learning, which reduce centralized data exposure by performing local training. However, recent studies indicate that achieving an optimal balance between privacy and utility in large-scale learning tasks remains highly challenging [10].

In conclusion, time and dynamic hazard have been recognized. Lacopini and some other people study the evolution of educational social systems, but Aziz emphasizes how to mitigate privacy risks regarding cross-platform identity association. Wu explains that cross-platform apps and platform sharing cause exposure of sensitive attributes. Thus, these studies need a unified systematic structure to rate hazards in different environments.

3 Methodology

3.1 Datasets and preprocessing

To guarantee the universality and comparability, this article uses multiple open social network datasets. Stanford Network Analysis Platform (SNAP) datasets (e.g., Facebook, Twitter, Google+) provide social graph structures and user attributes for attribute inference and link prediction tasks. Kaggle datasets contain social interaction records and user profiles, which are particularly useful for attribute inference and model inversion experiments. Location-based social networks, such as Gowalla and Foursquare, include spatiotemporal check-in data that enables temporal privacy analysis and cross-platform privacy research. Table 1 presents detailed information about the datasets used in this study.

Table 1. Dataset information.

Dataset	Users	Edges	Key Attributes
SNAP-Twitter	81,306	1,768,149	User attributes
SNAP-Google+	107,614	13,673,453	Circle relationships
Kaggle Social	~100,000	~1,000,000	Demographic data
Gowalla (SNAP)	196,591	950,327	6,442,890 locations
Foursquare (SNAP)	639,014	3,214,986	33,278,683 check-ins

These datasets are preprocessed to extract user attributes (e.g., gender, age, region), network structure features (e.g., degree distribution, clustering coefficient, community structure), and behavioral features (e.g., posting frequency, likes, reposts, mobility patterns).

The preprocessing steps include:

• Unifying node ID and deleting records that are missing in significant numbers.

• Extracting user attributes, like gender, age, and locations. Some sensitive attributes are better referred to.

• Calculating features of network structure, like frequency distribution, coefficient of clustering, and community structure, which are used for linkage predictions and attribute inference.

• Extracting behavior characteristics, like frequency of posts, retransmission, giving likes, or tracing the journey. These features provide supplementary information for attacking models.

• Cross-platform alignment: In cross-platform experiments, using matching strategies based on usernames and their interest similarity, to build different relationships among users on different platforms.

3.2 Experimental design

This research uses a novel approach that combines theoretical modeling and empirical experiments to develop a unified framework for quantifying privacy leaks. The framework includes four core modules:

Empirical Experiments: Describing the intensity of privacy leakage using experimental metrics such as entropy reduction, mutual information, and re-identification probability. These metrics not only quantify the degree of privacy leakage but also provide a standardized framework for comparing different attack strategies.

Attacking Environments: Simulating various attack strategies on real social network datasets, including attribute inference, link prediction, and model inversion. These environments replicate the adversaries' capabilities at different levels of information access.

Defense Mechanisms: Implementing strategies such as anonymity, differential privacy, hybrid defenses, and others, and comparing their effectiveness across different environments. Additionally, federated learning is introduced as a novel decentralized defense approach to evaluate risk differences compared to centralized data publishing models [11].

Experimental Evaluation: Conducting systematic analyses using metrics such as attack success rate, data utility retention, and privacy effectiveness. Further research explores the evolution of privacy risks in cross-platform and dynamic networks.

The experiment process follows these steps: choosing datasets and finishing pretreatment, implementing three types of attacks, recording attack mission success rates, and comparing protection effectiveness. Finally, counting and getting results, then conducting cross-platform and dynamic network analysis [12].

3.3 Theoretical framework

This study proposes a unified framework to quantify privacy leakage in social networks by integrating theoretical modeling with empirical experiments. Theoretically, privacy risks are quantified using mathematical indicators such as entropy, mutual information, and re-identification probability, which measure the extent to which adversaries can infer sensitive attributes. Empirically, multiple real-world social network datasets are employed to simulate attacks and assess the effectiveness of various defense mechanisms.

The framework consists of three components: Privacy Indicators, Attack Scenarios, and Defense Mechanisms. Privacy indicators define metrics to measure information leakage, including reductions in information entropy, accuracy of attribute inference, and success rates of link prediction. Attack scenarios implement representative adversarial models, including attribute inference, link prediction, and model inversion attacks. Defense mechanisms integrate baseline and enhanced privacy-preserving techniques and evaluate their effectiveness against a variety of attacks.

Three categories of privacy attacks are considered. Attribute Inference Attacks involve predicting sensitive attributes (e.g., gender, political preferences) based on graph and behavioral features. Link Prediction Attacks mean inferring hidden or sensitive connections within a social network, such as undisclosed friendships or relationships. Model Inversion Attacks involve recovering private information from trained graph neural networks, as demonstrated by GraphMI.

This article focuses on implementing and comparing several privacy-preserving mechanisms. Anonymization-based methods include k-anonymity, l-diversity, and structural perturbation techniques. Differential privacy mechanisms involve adding calibrated noise to query results or adjacency matrices. Hybrid defense combines differential privacy with graph structure preservation, aiming to maintain strong privacy protection while minimizing utility

loss. Federated learning approaches involve decentralized training methods that minimize the centralized exposure of raw data.

3.4 Evaluation metrics

The effectiveness of attack and defense mechanisms is evaluated using a comprehensive set of performance metrics, with particular attention to the Attack Success Rate of inference and prediction attacks. True Positive (TP) and True Negative (TN) represent correct predictions for positive and negative classes, respectively, while False Positive (FP) and False Negative (FN) represent prediction errors.

$$Accuracy=\left(\frac{TP+TN}{TP+TN+FP+FN}\right) \tag{1}$$

Measures the overall correctness of the classifier, i.e., the proportion of correctly predicted samples (both positives and negatives) among all samples. A higher value indicates better overall performance.

$$Precision=\left(\frac{TP}{TP+FP}\right) \tag{2}$$

Measures the proportion of true positives among all predicted positives. Reflects how many of the positive predictions are actually correct. A higher value means fewer false alarms (false positives).

$$Recall=\left(\frac{TP}{TP+FN}\right) \tag{3}$$

Measures the proportion of actual positives that are correctly identified. Reflects how many true positives are captured. A higher value means fewer misses (false negatives).

$$F1\text{-}Score=\left(\frac{2*Precision*Recall}{Precision+Recall}\right) \tag{4}$$

The harmonic mean of Precision and Recall, balancing the two metrics. A higher value indicates a better trade-off between precision and recall.

$$AUC=\int_0^1 TPR(FPR)d(FPR) \tag{5}$$

Where True Positive Rate (TPR) equals the number of TP divided by the total number of actual positive cases (the sum of True Positives and False Negatives, TP+FN). False Positive Rate (FPR) equals the number of FP divided by the total number of actual negative cases (the sum of False Positives and True Negatives, FP+TN).

Represents the area under the curve (AUC), measuring the classifier's ability to distinguish between classes across thresholds. A value closer to 1 indicates stronger discrimination capability.

Privacy Leakage Intensity is quantified using three complementary measures:

Entropy Reduction measures how much uncertainty about a sensitive attribute SSS is reduced after the adversary observes features XXX. A larger value indicates stronger privacy leakage.

$$\Delta H = -\sum s\, P(s) log P(s) - H(S|X) \tag{6}$$

Mutual Information captures the amount of shared information between sensitive attributes SSS and adversary's observed features XXX. A higher value means stronger correlation and more leakage.

$$I(S;X) = H(S) - H(S|X) = \sum P(s,x) * log \frac{P(s,x)}{P(s) * P(x)} \tag{7}$$

Re-identification Probability (Possibility of a user matching correctly) represents the probability that the adversary correctly matches an individual's identity. The overall average success rate reflects the severity of re-identification risk.

$$Preid = \frac{Number\ of\ corectly\ re-identified\ users}{Total\ users\ targeted} \tag{8}$$

Privacy Preservation Rate (Decreasing rate of defense strategies compared with attack success rate) measures how much the defense strategy reduces the attack success rate compared to the baseline (no defense). A higher value (closer to 1) indicates stronger privacy protection, while a negative value means the defense is counterproductive.

$$PPR = 1 - \left(\frac{ASR(defense)}{ASR(baseline)}\right) \tag{9}$$

In Data Utility Retention (Defense dataset can keep original function for attacks), U is a utility metric. It evaluates how much of the dataset's original utility is preserved after applying defenses. A value close to 1 means the defense maintains the dataset's usefulness for legitimate tasks while still protecting privacy.

$$UR = \frac{U(defense)}{U(baseline)} \tag{10}$$

Utility-Privacy Trade-off: An analysis of how defense methods balance privacy guarantees with the usability of models and data.

Temporal and Cross-Platform Risk Analysis: Comparing results across static versus dynamic networks, and single-platform versus multi-platform fusion.

4 Results and Analysis

This research implements three typical types of privacy attacks on multiple public social network datasets and introduces multiple defense mechanisms based on this basis. The experimental results not only verify notable privacy leak hazards in social networks, but also announce a balanced feature in different defense mechanisms and privacy. The paper

analyzes from three aspects: attacking effects, defense mechanism performance, and cross-platform or dynamic risks [13].

4.1 Attribute inference attacks

In evaluating attack effectiveness, the experiment first examines the performance of attribute inference attacks. Under the Twitter dataset, utilizing graph neural networks and user behavior characteristics to construct disaggregated models, in the case of no defense, about 82% accuracy can predict the gender of users, and a 74% accuracy rate can predict political tendency. This shows that although users don't need to announce sensitive information, user social relationships and interactions can be reliable features, and utilized by attackers.

For a further understanding, in the Kaggle user dataset, using models combined by Extreme Gradient Boosting (XGBoost) and Graph Sample and Aggregate (GraphSAGE), the AUC of predicted gender is 0.86, showing strong applicability under attack in different data source conditions. Overall, attribute inference attacks always have higher effectiveness in different platforms, which verifies privacy sensitivity between social structure and behavior features. Table 2 shows the performance of attribute inference attacks across different datasets.

Table 2. Attribute inference attack results (%).

Dataset	gender	age	political orientation	Average AUC
Twitter	82.1	71.4	74.3	0.86
Kaggle	84.5	73.2	75.1	0.88
Google+	79.6	69.8	72.5	9.84

4.2 Link prediction attacks

In the experiment of link prediction attacks, research focuses on the recovery of unpublic relationships. In the Facebook dataset, Node2Vec was employed to create embedded vectors, and prediction was achieved by combining logistic regression. The AUC is 0.88, and precision@10 still keeps a size of 0.73. This result shows that potential relationships of social networks can be referred to easily. Although users can hide some parts of relationships, the neighborhood structure based on graph neural networks is more prominent. It has 0.91 of AUC in forecasting tasks under privacy sensitivities.

These results show that link prediction not only announces explicit social structure, but also uncovers relationships that users are deliberately hiding, which raises some serious privacy concerns. Table 3 presents the results of link prediction attacks using different methods.

Table 3. Link prediction attack results (AUC).

Dataset	Deepwalk	Node2Vec	GNN-based	Average
Facebook	0.82	0.85	0.88	0.85
Google+	0.87	0.89	0.91	0.89
Twitter	0.81	0.83	0.86	0.83

4.3 Model inversion attacks

Model inversion attacks verify situations where attackers cannot directly access original datasets, but can still recover sensitive information through trained models. In the Kaggle dataset, GraphMI structure can recover users' interest tags with 71% accuracy, and the

entropy of the recovery results is about 22% less than the true distribution. This shows that model parameters contain strong privacy signals.

In the Twitter dataset, the performance of model inversion is relatively weak, but 65% accuracy can be reached. Research results show that privacy risk can't be eliminated by only relying on model parameters or shared data on output, and privacy leaks can get worse due to models having strong expressive power. Table 4 summarizes the model inversion attack results.

Table 4. Model inversion attack results.

Dataset	Sensitive Attribute	Inference Accuracy	Entropy Reduction
Kaggle	User Interest	71.2	22%
Twitter	Profile Location	65.3	18%

4.4 Defense mechanisms evaluation

In the evaluation experiment of defense mechanisms, this research found that distinct differences exist in the effects between decreasing attack success mission rates and keeping data utility. Anonymity methods use attribute reduction to get an accuracy decrease of 30% in k=10, but linkage predictions show that AUC decreases from 0.85 to 0.67. This shows that structural perturbation has a significant impact on graph mining.

Differential privacy shows that success mission rates decrease 50% when ε=1, but efficacy gets worse, and some parts of task performance decrease by over 25%. By comparison, mixed defense shows a better balance. When mild structural disturbance is combined with medium intensity differential privacy (ε=5), attack mission success rates decrease by about 40%, and effectiveness remains at 80%. Federated learning has outstanding performance in preventing model inversion attacks, and recovery accuracy decreases by more than 20%. However, large-scale training incurs costs for communication and computing. Table 5 compares the effectiveness of different defense mechanisms.

Table 5. Defense mechanisms comparison.

Defense mechanisms	Decrease of attack success mission rate	Keep data useful percentages	Other information
Anonymization(k=10)	-30%	65%	Loss of network structure
Differential Privacy (ε=1)	-50%	60%	Data efficiency decreases
Mixed Defense	-40%	80%	Balance between privacy and effect
Federal study	-20%	85%	Expensive communication

4.5 Temporal and cross-platform risks

In analyzing cross-platform and dynamic risks, the experiment announces the amplification effect of privacy leaks. After combining Twitter with Foursquare, the accuracy of gender judgment increases from 76% to 92%, which means cross-platform enhances the ability of attackers. In a time series analysis of the Gowalla dataset, the recognition probability in 6 months increases by about 15%, which means privacy hazards progressively accumulate with time. Table 6 shows the temporal and cross-platform risk analysis results.

Table 6. Temporal and cross-platform risk analysis.

Environment	Accuracy of baseline	Accuracy after mixed	Increasing hazard
Twitter	76%	-	-
Twitter+Foursquare	-	92%	+16%
Gowalla(0-3 months)	65%	-	-
Gowalla(6 months)	80%	-	+15%

In conclusion, social networks have enough privacy risks in different aspects, like attribute inference, linkage predictions, and model inversion. Anonymous and differential privacy provide enough protection associated with utility loss, and mixed defense has better performance in balance. Federated learning shows a newly developing latent ability. Cross-platform and dynamic scenes increase privacy risks and provide important inspiration for social platform governance.

5 Conclusion

This research constructs a uniform structure of social network privacy leak analysis, and rates the attribute reference, linkage prediction, and model inversion in multiple real datasets. The result of experiments says the social network has a distinct privacy leak hazard. Although in an anonymous situation, attackers can still refer to sensitive information through the network structure and behavioral characteristics. Different defense mechanisms have different advantages and disadvantages: Anonymous and differential privacy will decrease the attack success rate, but they always have a larger utility loss. Mixed loss has a good balance between privacy protection and data availability. The federal study has a good performance in defensive model inversion attack, but still has some problems in large-scale applications like communication and counting. For a further study, cross-platform binding and dynamic Internet greatly amplify privacy risk. Overall, this research not only fills the blank space about privacy leakage quantification analysis, but also gives some guidance for social platforms and monitoring organizations about privacy protection and data usage. In the future, the research can focus on defense distribution and cross-domain risk governance, which are used for sustainable projects. Something that needs to be emphasized is that the privacy leak of social networks shows more complex and dynamic features with the development of artificial intelligence and data technology. This article gives the structure and results of empirical evidence, not only giving new research ideas for the academic world, but also providing a scientific basis for industry practice and policy-making. In the future, social networks will become safe and have sustainable development only in this situation: the development of technical innovation, policy making, and user awareness.

References

1. Q. Yuan, Z. Zhang, L. Du, M. Chen, P. Cheng, M. Sun, PrivGraph: Differentially private graph data publication by exploiting community information, in Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), (2023), pp. 3241-3258
2. G. Beigi, H. Li, A survey on privacy in social media: Identification, mitigation, and applications. ACM Computing Surveys **53**, 1-38 (2020)
3. I. Iacopini, T. Milojević, V. Latora, G. Petri, The temporal dynamics of group interactions in higher-education social systems. Nature Communications **15**, 2345 (2024)

4. Y. Feng, J. Zang, X. Sun, B. Li, Leakage and privacy at inference time: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 9981-10002 (2022)
5. I. Iacopini, T. Milojević, V. Latora, G. Petri, The temporal dynamics of group interactions in higher-order social networks. Nature Communications **15**, 2346 (2024)
6. T.T. Mueller, C.M. O'Keefe, Differentially private guarantees for analytics and machine learning on graphs: A survey of results. Journal of Privacy and Confidentiality **14**, 1-35 (2024)
7. G.M. Garrido, C. Laplace, J. Soria-Comas, J. Domingo-Ferrer, D. Rebollo-Monedero, Lessons learned: Surveying the practicality of differential privacy in the industry. Proceedings on Privacy Enhancing Technologies **2023**, 1-22 (2023)
8. X. Han, Y. Wang, C. Zhang, Q. Yang, Privacy-preserving network embedding against private link inference attacks. IEEE Transactions on Dependable and Secure Computing **21**, 567-582 (2024)
9. Y. Li, K. Zhao, W. Chen, T. Zhou, Differentially private knowledge graph neural networks for recommender systems. Knowledge-Based Systems **297**, 110987 (2025)
10. Z. Xu, H. Lin, J. Wang, C. Yu, Enhancing federated learning-based social recommendation with graph attention networks. Neurocomputing **600**, 128-140 (2025)
11. X. Ma, H. Chen, Y. Zhang, A differential privacy protection algorithm for network sensitive information based on SVD. Scientific Reports **13**, 98765 (2023)
12. I. Iacopini, T. Milojević, V. Latora, G. Petri, The temporal dynamics of group interactions in higher-order social networks. Nature Communications **15**, 2346 (2024)
13. Z. Nie, J. Smith, A. Brown, A framework for the design of privacy-preserving record linkage systems. Privacy **5**, 45-62 (2025)