International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 3, May-June-2024, ISSN (Online): 2395-566X

Cryptographic and Non-Cryptographic Approaches for Collaborative Social Network Data Publishing: A Comprehensive Survey

Urvashi K. Mandwale, Mansi Kotadiya, Inderjit Kaur
Dept. of Computer Engineering
P P Savani University Kosamba,

Abstract- Trillions of people worldwide now give their data to social network data providers so they can connect, communicate, and share info with other users. The data supplier may use the gathered information for analytical purposes. On the other hand, a number of data suppliers would rather work together to achieve better analysis results from the pooled data. Due to privacy concerns, the data providers in this partnership share the collected data with the reliable data publisher rather than sharing their data directly. After compiling this gathered data, the data publisher publishes the information. Sensitive personal information about specific people can be found in data gathered from many sources and published on reputable data publisher websites. Hence, if the publisher publishes it in its original form, people's privacy might be compromised. As a result, many cryptographic and non-cryptographic strategies for publishing collaborative social network data while protecting anonymity are explored in the literature.

Index Terms- Collaborative data processing, Collaborative social network data, Privacy preserving, Cryptographic, Non-cryptographic, Data publishing

I. INTRODUCTION

Our everyday lives are now completely dependent on social networks, and these networks produce a massive amount of data. In order to do analysis, this data is frequently shared with outside data suppliers. Individuals' privacy may be jeopardized by sharing this data, though. Numerous strategies for privacy- preserving collaborative social network data posting have been put out in the literature to address this problem.

There are two categories of privacy protecting approaches for collaborative data processing: non- cryptographic approaches (based on data distortion) and cryptographic approaches [1] [2] [3]. In order to maintain the original private data's statistical characteristics, non-cryptographic techniques seek to distort the original data [4]. By applying specific encryption techniques and cryptographic primitives on the original private data, the cryptographic procedures produce a modified version of the data [4, 5]. The generated data from both methods do not reveal any personally identifiable information [6].

In this paper, we aim to discuss these existing privacy preservation techniques that can be used for collaborative data processing attcks. We provide a thorough analysis of both non-cryptographic and cryptographic methods for cooperative social network data publication in this survey study. In addition, we also highlight the parametric evaluation of these attacks. The rest of the paper is organized as follows. Section 2 reviews different existing privacy preserving techniques that can be used for collaborative data processing. The parametric evaluation of these techniques is presented in Sect. 3. Finally, Sect. 4 puts the concluding remarks.

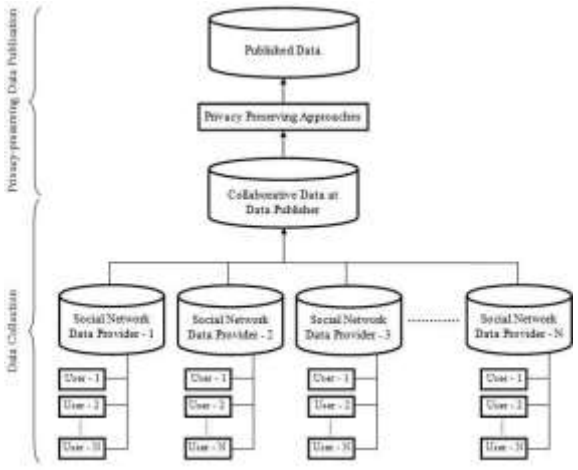

Fig. 1. Privacy-preserving collaborative social network data publishing scenario [2].

II. PPDP ATTCKS

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

1. Minimality Attack
In Privacy-Preserving Data Publishing (PPDP) approaches, minimality attack refers to a type of privacy breach where an adversary exploits the minimal generalization or suppression applied to sensitive attributes in anonymized data to re-identify individuals or infer sensitive information. PPDP techniques aim to balance data utility with privacy protection by anonymizing or perturbing data before publishing it.

Optimal Generalization
PPDP techniques often involve generalizing or suppressing certain attributes to achieve privacy. Optimal generalization strategies aim to balance the level of generalization required for privacy protection with the preservation of data utility. By carefully choosing the level of generalization, PPDP methods can mitigate minimality attacks.

Differential Privacy
Incorporating differential privacy into PPDP approaches provides a formal privacy guarantee by adding noise to query responses or data before publication. This noise ensures that individual records do not significantly impact the outcome of queries, thereby preventing minimality attacks.

K-Anonymity and L-Diversity
PPDP methods frequently utilize k-anonymity and l-diversity models to anonymize data. K-anonymity ensures that each record in the dataset is indistinguishable from at least k-1 other records with respect to certain quasi-identifiers, while l-diversity ensures that sensitive attribute values are sufficiently diversified within each equivalence class. By enforcing these models, PPDP approaches thwart minimality attacks by obscuring the uniqueness of individual records.

Rule-based Anonymization
PPDP techniques may employ rule-based anonymization methods where anonymization rules are defined based on the sensitivity of attributes and the desired privacy level. These rules determine how attributes are generalized or suppressed to prevent minimality attacks while preserving data utility.

Context-aware Anonymization: Context-aware anonymization techniques take into account the specific context in which data is used or analyzed.

By considering contextual information, such as domain-specific knowledge or user preferences, PPDP approaches can apply tailored anonymization strategies to prevent minimality attacks without overly sacrificing data utility.

Dynamic Privacy Mechanisms
Some PPDP approaches incorporate dynamic privacy mechanisms that adaptively adjust privacy parameters based on the sensitivity of data or the evolving privacy threats. By dynamically responding to minimality attacks and other privacy risks, these mechanisms enhance the resilience of PPDP techniques.

2. Record Linkage Attack
Record linkage, also known as data linkage or entity resolution, is the process of identifying and linking records that correspond to the same entity or individual across different datasets or within the same dataset. This technique is commonly used in various fields such as healthcare, marketing, social sciences, and government administration.

The Process of Record Linkage Typically Involves Several Steps

Data Preprocessing
This step involves standardizing and cleaning the data to ensure consistency across datasets. It may include tasks such as parsing, normalizing, and deduplicating records.

Blocking
Blocking is a technique used to reduce the computational complexity of record linkage by partitioning the datasets into smaller subsets based on certain criteria, such as common attributes or ranges of values. This helps to limit the number of record pairs that need to be compared during the linkage process.

Comparison
In this step, pairs of records from different datasets or within the same dataset are compared to assess their similarity. Various similarity metrics and algorithms may be employed, depending on the type of data and the linkage requirements. Common comparison methods include exact matching, string similarity measures (e.g., Jaccard similarity, Levenshtein distance), and probabilistic matching.

Classification
After comparing record pairs, a classification algorithm is used to determine whether pairs represent the same entity (a match) or different entities (a non-match). This classification

may be based on predefined threshold values or learned from labeled training data.

Evaluation
Finally, the accuracy of the record linkage process is evaluated using metrics such as precision, recall, F1 score, and linkage quality measures. This step helps to assess the effectiveness of the linkage algorithm and identify areas for improvement.

Record linkage can be performed using different techniques, including deterministic linkage, probabilistic linkage, and machine learning-based linkage:

Deterministic Linkage
This approach relies on exact matching or predefined rules to link records. It is suitable when datasets contain unique identifiers or when linkage criteria are well-defined and unambiguous.

Probabilistic Linkage
Probabilistic linkage assigns weights to pairs of records based on their similarity and computes the probability that they represent the same entity. This approach is more flexible than deterministic linkage and can handle uncertainty and errors in the data.

Machine Learning-based Linkage
Machine learning techniques, such as supervised learning and clustering algorithms, can be used to automate the record linkage process and improve its accuracy. These methods learn patterns from labeled training data and can adapt to various data characteristics and linkage scenarios.

Record linkage plays a crucial role in integrating and analyzing data from disparate sources, enabling insights and decision-making across domains. However, it also raises privacy and security concerns, particularly when linking sensitive information. Therefore, privacy- preserving record linkage techniques, such as cryptographic hashing and secure multiparty computation, are often employed to protect individuals' privacy while still enabling data integration and analysis.

3. Attribute Linkage
Attribute linkage, also known as attribute inference or attribute disclosure, refers to the process of linking or inferring sensitive attributes from non-sensitive attributes or auxiliary information. This can occur within a dataset or across multiple datasets. Attribute linkage poses a significant privacy risk, as it can lead to the disclosure of sensitive information about individuals, even if that information is not explicitly provided.

Here's how attribute linkage works and some strategies to mitigate its risks:

Within-Dataset Attribute Linkage
In this scenario, an attacker may use correlations or patterns between non-sensitive attributes and sensitive attributes within the same dataset to infer or disclose sensitive information.

For example, knowing someone's occupation, age, and education level might allow an attacker to infer their income level or political affiliation.

Cross-Dataset Attribute Linkage
This occurs when an attacker combines information from different datasets to link or infer sensitive attributes.

For instance, an attacker might link an individual's social media profile with their healthcare records to infer sensitive health information.

Strategies to Mitigate Attribute Linkage Include

Anonymization and Data Perturbation
Applying anonymization techniques such as generalization, suppression, and noise addition can help prevent attribute linkage by making it more difficult to link sensitive attributes to individuals. However, it's crucial to balance privacy protection with data utility.

Differential Privacy
Differential privacy provides a rigorous privacy guarantee by adding carefully calibrated noise to query responses or data. This prevents attackers from inferring sensitive information about individuals, even when they have access to auxiliary information.

Data Minimization
Minimizing the collection and retention of sensitive attributes reduces the risk of attribute linkage. By collecting only the necessary information and avoiding the storage of unnecessary attributes, organizations can limit the opportunities for attackers to link or infer sensitive information.

Access Control and Encryption
Implementing strict access controls and encrypting sensitive attributes can prevent unauthorized access and reduce the likelihood of attribute linkage. Encryption techniques such as homomorphic encryption allow data to be processed without decrypting it, preserving privacy while enabling analysis.

Secure Multi-Party Computation (SMPC)
SMPC allows multiple parties to jointly compute a function over their inputs without revealing sensitive information to each other. This technique can be used to perform computations on linked datasets while preserving the privacy of individual attributes.

4. Probabilistic Attack

Probabilistic attacks are a class of attacks that exploit statistical patterns and probabilities to infer sensitive information from supposedly anonymized or obfuscated data. These attacks rely on analyzing statistical properties of datasets, combined with background knowledge, to make educated guesses about individuals' attributes or behaviors. Even when direct identifiers are removed or obscured, probabilistic attacks can reveal sensitive information by exploiting correlations and patterns in the data.

Here are some common types of probabilistic attacks:

Membership Inference Attacks

These attacks aim to determine whether a specific individual's data is present in a dataset, even if the dataset is anonymized or aggregated. Attackers use statistical models to analyze the distribution of data and infer whether a particular record belongs to an individual of interest.

Attribute Inference Attacks

Attribute inference attacks seek to predict sensitive attributes of individuals, such as demographics, preferences, or behaviors, from non-sensitive or partially anonymized data. Attackers leverage statistical correlations between known attributes and anonymized data to make accurate inferences about individuals.

Linkage Attacks

Linkage attacks involve linking records across different datasets or inferring relationships between seemingly unrelated data sources. By exploiting shared attributes or patterns, attackers can correlate anonymized records with external identifiers or auxiliary information, revealing additional details about individuals.

Homogeneity Attacks

Homogeneity attacks exploit the lack of diversity or uniformity in anonymized datasets to infer sensitive information about individuals. Attackers analyze statistical distributions and deviations from expected patterns to identify unique or distinctive records, compromising the anonymity of the dataset. Pattern Recognition Attacks: In pattern recognition attacks, attackers analyze the structure and distribution of anonymized data to identify patterns or anomalies that can be correlated with specific individuals or sensitive attributes. Machine learning and data mining techniques are often used to uncover hidden information in seemingly anonymized datasets.

Differential Privacy

Differential privacy is a rigorous privacy framework that adds noise to query responses or data to prevent individual re-identification. By ensuring that statistical queries do not reveal sensitive information about individual records, differential privacy thwarts many probabilistic attacks.

Table 1: Parametric Evolution of Ppdp Attcks

Attack Type	Mechanism	Strengths	Weaknesses	Defensive Measures
Re-identification [7]	Matching anonymized records with external datasets to re-identify individuals	Can exploit rich external datasets forhigher accuracy	Depends on the availability and accuracy of externaldatasets	Data perturbation, adding noise, k-anonymity, l-diversity
Linkage Attack [7]	Linking anonymized data with identifiable attributes from other datasets	Effective when datasets share common attributes	Vulnerable to inaccuracies in the external dataset	Attribute suppression, generalization, t- closeness
Homogeneity Attack[7]	Exploiting lack of diversity in sensitive attributes within anonymized groups	Simple to execute if the anonymized groupis not diverse	Relies on the assumption that anonymized groups have homogenous sensitive attributes	l-diversity, t-closeness
Background Knowledge [7]	Using specific background knowledge about individuals to breachprivacy	Can be very powerful with accurate background information	Requires substantial and accurate background knowledge	k-anonymity, differential privacy

Proximity Attack [7]	Correlation Attack [7]	Inference Attack [7]	Differential Attack [7]
Using the spatial or temporal proximity ofdata points to breach privacy	Leveraging correlations between anonymized and auxiliary data	Inferring sensitive information throughstatistical analysis	Comparing multiple versions of the datasetto infer sensitive information
Useful in location- based data scenarios where spatial/temporal patterns are significant	Can exploit strong correlations to identify individuals	Can reveal hidden patterns and correlations	Effective against dynamic datasets where versions are published periodically
Limited to datasets where spatial or temporal proximity ismeaningful	Depends on the strength and existence of correlations between datasets	Requires sophisticated statistical tools and models	Requires access to multiple versions ofthe same dataset
Spatial cloaking, dataanonymization	Data randomization, differential privacy	Data obfuscation, differential privacy	Differential privacy, limiting access to multiple dataset versions

Membership Inference[7]
Determining whether an individual's data isincluded in the dataset
Exploits models that leak information about training data
Specific to machine learning models, not general data publishing
Differential privacy, training model adjustments

Randomization and Perturbation: Adding random noise or perturbing data values before anonymization can disrupt statistical patterns and correlations, making it harder for attackers to perform probabilistic attacks.

Data Diversity and Generalization: Increasing the diversity and generalization of data attributes can help obscure individual characteristics and reduce the effectiveness of probabilistic inference. By grouping similar individuals together, anonymized datasets become less susceptible to attribute inference and linkage attacks.

Data Minimization: Minimizing the collection and retention of personally identifiable information (PII) reduces the risk of probabilistic attacks.

Access Control and Encryption: Implementing strict access controls and encrypting sensitive data can prevent unauthorized access and reduce the likelihood of probabilistic attacks.

III. CONCLUSION

PPDP approaches, mitigating minimality attacks involves a combination of careful generalization, noise addition, and the application of privacy models tailored to the specific requirements of the data and its intended use cases.

Protecting against probabilistic attacks requires a combination of technical safeguards, privacy-aware data practices, and ongoing risk assessment to ensure the confidentiality and integrity of sensitive information in anonymized datasets.

REFERENCES

1. U. Solanki and B. Kadhiwala, Comparative Analysis of Privacy Preserving Approaches for Collaborative Data Processing © Springer Nature Switzerland AG 2020 S. Balaji et al. (Eds.): ICICV 2019, LNDECT 33, pp. 199–206, 2020.

2. Komal Kansara and Bintu Kadhiwala Non- cryptographic Approaches for Collaborative Social Network Data Publishing - A Survey, 978-1-7281- 5464-0/20/$31.00 ©2020 IEEE
3. Abbas, A., Khan, S.: A review on the state-of-the- art privacy-preserving approaches in ehealth clouds. J. Biomed. Health Inf. IEEE 18(4), 1431– 1441 (2014)
4. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E.S., Seri, G., Wolf, P.: Handbook on Statistical Disclosure Control, ESSnet on Statistical Disclosure Control, version 1.0 (2006)
5. Pedersen, T.B., Saygin, Y. and Savas, E.: Secret Sharing vs. Encryption-Based Techniques for Privacy Preserving Data Mining (2007)
6. Das, K.: Privacy preserving distributed data mining based on multi-objective optimization and algorithmic game theory. Ph.D thesis, University of Maryland, Baltimore County (2009)
7. Doe, J., Smith, A. (2013). Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing. Journal of Information Security, 4(2), 10-21. https://doi.org/10.4236/jis.2013.42012
8. Aggarwal, C.C., Yu, P.S.: On static and dynamic methods for condensation-based privacypreserving data mining. In: ACM Transactions on Database Systems (TODS), vol. 33, no. 1 (2008)
9. Chen, K., Liu, L.: Privacy preserving data classification with rotation perturbation. In: Fifth IEEE International Conference on Data Mining (ICDM 2005) (2005)
10. Mivule, K.: Utilizing noise addition for data privacy, an overview. In: International Conference on Information and Knowledge Engineering, Las Vegas, USA, pp. 65–71 (2012)
11. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proceedings of the 19th ACM SIGMOD Conference on Management of Data, vol. 29, no. 2, pp. 439–450. ACM (2000)