International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I12P112

Volume 73 Issue 12, 151-161, December 2025

Original Article

Neighborhood-Based Similarity Anonymization (NSA): A Multi-Level Approach for Graph Anonymization

Mariam RAMDI[1*], Ouafae BAIDA[1], Abdelouahid LYHYAOUI[1]

[1]*LTI Lab, ENSA of Tangier, Abdelmalek Essaâdi University, Tangier, 90000, Morocco.*

**Corresponding Author: mariam.ramdi@etu.uae.ac.ma*

Received: 04 August 2025 Revised: 19 November 2025 Accepted: 21 November 2025 Published: 19 December 2025

Abstract - *Social network emergence has enabled the dissemination of vast amounts of data, beneficial to numerous applications but detrimental to privacy. Typical anonymization approaches often face the challenge of finding a balance between the utility and the protection of privacy, leading to excessive information loss or weak anonymization. In this paper, a new methodology, Neighborhood-Based Similarity Anonymization (NSA), is proposed, which strengthens privacy through an evaluation of user similarity at multi-level network neighborhoods. Unlike conventional approaches that consider direct user associations, NSA adopts 1-hop (direct), 2-hop (friends-of-friends), and 3-hop (third-degree) neighborhood similarities for intelligent edge elimination decisions aimed at retaining the connected graph, where 'hop' defines the distance between users in the network graph. With the real-world Twitter dataset, the efficiency of the proposed method, NSA, for the protection of privacy with retention of structural integrity, is shown to outperform common similarity-based anonymization techniques with an outstanding balance between privacy and utility.*

Keywords - *Privacy, Reidentification, Data Utility, Anonymization, Social Networks.*

1. Introduction

The rise of social networks has resulted in substantial quantities of patterns and impact configurations. Consequently, successful data accessibility now facilitates numerous anonymization solutions that must safeguard user privacy without altering the network's topology or burdening the diverse applications, particularly in social behavior analysis, while ensuring the network remains intact for analytical purposes. Suggestions. Social network data is a valuable resource for academic research and standard anonymization techniques for commercial usage, since it discloses user intentions, including edge modification, edge switching, and node association patterns. Nevertheless, the escalating use of social network data, together with evolving connection dynamics and impact, amplifies the danger to users' similarity metrics or closeness, a significant approach to this privacy. Information acquired from social networking platforms might disclose sensitive personal details. This category pertains to user similarity-based anonymization methods that reveal sensitive personal information, as outlined in [1]. It emphasizes direct connections for benign purposes, highlighting the need for effective user protection to preserve the analytical value of the data.

The major issue with anonymization of social networks is hitting a balance between the protection of privacy (protection from reidentification) and the preservation of data utility (preservation of analytical usefulness). Conventional methods of anonymization, including k-anonymity, l-diversity, and differential privacy, mostly use tabular or relational representations. When applied to social network data, they may result in significant information loss or fail to maintain the relationship structure necessary for successful analysis. Anonymization for social networks considers the intrinsic complexity of the network, including both direct relationships (e.g., friendships or follower connections) and indirect relationships that influence the underlying patterns and structures inside the network. Consequently, an efficient anonymization technique must safeguard user privacy while minimizing changes to the network's topological structure and maintaining analytical value. Standard approaches to social network anonymization include edge modification, edge swapping, and node clustering, often achieved by grouping users based on similarity metrics or proximity. A notable method in this category is the user similarity-based anonymization approach described in [1], which focuses on direct connections to anonymize user data by selectively removing or modifying edges. While this approach offers privacy protection by altering identifiable connections, it is limited in its ability to capture broader relational patterns inherent in social networks. As a result, privacy breaches remain possible when adversaries exploit indirect relationships such as second or third-degree connections to re-identify anonymized nodes.

While anonymization in social networks has been extensively studied, most current approaches remain dependent on direct (1-hop) connections and neglect privacy risks from indirect or multihop connections. This leaves a research gap on how to fully protect users from reidentification while preserving the network's analytical utility. Traditional approaches like k-anonymity, l-diversity, and differential privacy distort the relational structure when applied to graph data, resulting in the loss of critical topological information. Likewise, current similarity-based methods, such as [1], only focus on local connections and ignore the effect of higher-order relations, which can still expose sensitive user patterns through indirect associations

This paper presents the Neighborhood-Based Similarity Anonymization (NSA), which is an all-encompassing anonymization framework aimed at addressing the limitations of classical approaches through the reinforcement of privacy protections at varying neighborhood levels. NSA applies 1-hop, 2-hop, and 3-hop neighborhood similarities rather than solely relying on 1-hop (direct) relations. The resulting methodology is an advance over a more sophisticated anonymization framework that takes cognizance of both the immediate and the extended relational circumstances. Through multi-level anonymization of the network, the NSA not only ensures protection for privacy but also mitigates the risk inherent in indirect relations. The structural completeness of the network, which is invaluable for conducting network analysis, is also maintained through this method.

1-hop Similarity describes direct relations, 2-hop Similarity describes relations through an intermediary user, and 3-hop Similarity describes relations split through two intermediaries. The multi-level neighborhood analysis permits the NSA to obscure user identities with retention of pertinent network properties, for example, clustering and connectivity, pertinent to downstream applications. The goal of the NSA is to preserve privacy to the greatest extent possible without inducing too many structural changes. They modify only those edges that are too similar at these multi-levels. The major contributions that this paper provides are:

- A novel multi-level neighborhood-based anonymization method called NSA that preserves indirect relations up to 3 hops. It provides stronger privacy protection than the classical single-level approaches.
- Using real social network data from Twitter to study the effects of the NSA on the architecture and connectivity of networks. This suggests that the NSA reduces vital properties, including average path length and clustering, that are useful for the study of networks.
- Setting right, the assessment of the efficacy of the NSA compared to other anonymization approaches sheds light on its higher potential for privacy protection, with the further benefit of achieving an equilibrium between utility and privacy needs.

The remainder of this paper is organized as follows: Section 2 reviews the related work on social network anonymization techniques and discusses their limitations. Section 3 outlines the main attack scenarios and presents the defense mechanisms adopted by the proposed NSA framework. Section 4 details the methodology and implementation of the proposed approach. Section 5 describes the experimental setup, evaluation metrics, and comparison with baseline methods. Section 6 presents and analyzes the results obtained from the experiments. Section 7 provides an in-depth discussion of the findings, and finally, Section 8 concludes the paper and suggests directions for future research.

2. Related Work

Anonymizing social networks has attracted wide interest among researchers, due to the wide challenge of bringing privacy to the users without losing the valuable elements of the data. A wide variety of techniques have been introduced for dealing with these conflicting objectives, each with some pros and cons. Here, some common categories of anonymization techniques for social network data are introduced.

Traditional anonymization techniques, i.e., K-anonymity, L-diversity, and T-closeness, are fundamental approaches designed for tabular datasets to protect privacy through the consolidation of alike persons, thus impeding the process of reidentification [2-4]. However, the application of these approaches to social network data could give rise to the loss of relational knowledge due to the weakness found in sustaining the highly intertwined, networked nature of the information. The main issue is to balance the utility of the data with the concern for privacy when anonymizing relations, where social network data often relies on indirect connections that are highly analytically valuable.

Differential privacy provides a firm mathematical framework for ensuring privacy through the addition of controlled noise to the dataset [5]. When applied to social networks, techniques built on differential privacy add noise to the edges or modify graph attributes to mask individual identities. However, this approach could interfere with critical network properties, such as clustering coefficients and the community structure, thus limiting its applicability for applications that require the retention of network topology. In addition, the computational cost of differential privacy creates formidable scalability challenges when working with large social network datasets [6, 7].

Similarity-based approaches anonymize social graphs by deleting or modifying edges under user similarity metrics. Similarity-based approaches try to find a privacy-utility balance with the preservation of useful relations among similar nodes. However, similarity-based approaches are likely to involve direct, 1-hop relations, and are hence prone

to privacy violation with the help of indirect relations (2-hop or 3-hop relations) [8]. The limitation prohibits the applicability of similarity-based anonymization to the retention of pervasive relational patterns prevailing among social graphs [2].

Clustering approaches also guarantee user privacy through the clustering of similar nodes within a network, with individual identities concealed within these aggregates. Approaches such as k-degree anonymity, where the nodes are clustered through similar degrees, could hide personal identities without losing the structural properties of the community networks [9]. Such clustering approaches, however, are also found to be highly inefficient under dynamic networks, where the structural configurations of the community networks are subject to temporal changes. For this reason, innovations like dynamic clustering were proposed; however, scalability and computational complexity issues remain significant areas of concern [10].

Multilayer neighborhood-oriented approaches answer the limitations of single-layer strategies through the anonymization of social networks that involve both direct and indirect ties, practically identifying patterns through 1-hop, 2-hop, and 3-hop linkages [8, 11]. The method promises enhanced protection for privacy from attacks that seek to use indirect connections for the purpose of reidentification. Recursive anonymization approaches follow suit through the repeated anonymization of the network through several levels; however, these approaches are likely to face issues with scalability, particularly in large-scale networks [9, 12].

Recently, the literature on graph anonymization has moved in two complementary directions: approaches with formal guarantees of graph confidentiality (variants of differential graph confidentiality) and more powerful reidentification attacks and methods based on embeddings and machine learning models (GNNs/matching) that leverage multihop similarities. Recent reviews indicate that differential confidentiality-based approaches maintain theoretical robustness but suffer from utility problems (modification of topological properties) and scalability issues for large networks. At the same time, recent work demonstrates learned de-anonymization attacks (neural/embedding-based) and membership inference attacks on graph models, validating the fact that considering multihop (2-hop, 3-hop) similarities is necessary to mitigate information leakage vectors. These results highlight the value of our NSA approach: by combining multi-level anonymization (1, 2, and 3-hop) and connectivity checks, NSA provides an interesting compromise between preserving utility and mitigating the risks highlighted by recent work [13].

All these approaches are compromises between structural properties for effective network analysis and privacy protection. The Neighborhood - Based Similarity Anonymization (NSA) method presented here tries to fill these gaps by anonymizing the network at macro and micro levels to enable direct and indirect relations between users. By incorporating 1-hop, 2-hop, and 3-hop neighborhood similarities, NSA provides a holistic approach to enabling privacy protection with the required structural properties for network analysis.

3. Problem Statement

3.1. Ethical Considerations in Social Network Data Use

The use of social media data raises ethical concerns about privacy, compliance, and the potential for indirect reidentification. Even if the data is not explicitly personal, relationships, interaction patterns, and textual similarities can expose preferences, sensitive behaviors, or enable identities to be inferred from unique patterns. These risks are compounded in social graphs, where multihop links represent personal relationships that attackers can leverage. Anonymization must therefore comply with data minimization, purpose limitation, and defense against structural attacks, and ensure the ethical use of anonymized data in a legitimate scientific context.

3.2. Technical Problem: Privacy–Utility Trade-Off in Graph Anonymization

While several works deal with anonymizing social networks, existing approaches still fail to provide sufficient protection against reidentification attacks while preserving their analytical utility. Classical anonymization techniques like k-anonymity, l-diversity, or differential anonymization are designed for tabular data and are not easily applicable to graph structures, as they can distort topological information. Moreover, similarity-based anonymization methods, which only consider direct (1-hop) relationships, only anonymize direct connections, leaving users exposed to 2-hop and 3-hop data. This edge-only focus restricts the capability to avoid reidentification with multi-level structured cues. Furthermore, global edge removal strategies tend to break graph connectivity, making the anonymized network less useful for downstream analysis.

Existing methods suffer from three limitations. First, similarity-based anonymization techniques only consider direct (1-hop) links, ignoring indirect similarities that attackers can leverage in multi-level attacks. Second, no current method ensures systematic anonymization of two- or three-hop neighborhoods while preserving structure, leaving high-level patterns vulnerable. Third, approaches that typically delete edges or add noise often compromise connectivity, extend paths, or break node clustering, rendering them less useful for analysis.

The proposed method, Neighborhood Similarity-Based Anonymization (NSA), directly addresses these limitations. The NSA employs a multi-level similarity assessment (1-hop, 2-hop, and 3-hop) to identify both direct and indirect relational patterns. It lowers multihop Similarity, which is a known way

for attackers to guess and re-identify structures. NSA also makes sure that targeted edge removal is done based on similarity thresholds and that graph connectivity is kept through connectivity preservation checks. Ultimately, the NSA has a flexible anonymization system that allows users to choose the level of anonymization (1-hop, 2-hop, or 3-hop) to meet varying privacy and utility requirements. This formulation clearly demonstrates that NSA is a method that addresses a specific research gap and enhances graph anonymization.

4. Novelty and Contribution

Neighborhood-Based Similarity Anonymization (NSA) is a novel approach that leverages multilevel neighborhood similarities to achieve anonymity. Unlike traditional approaches that restrict themselves to direct (1-hop) relationships or perform global edge removal, NSA calculates similarities up to the 3-hop, capturing both direct and indirect relational patterns. This multilevel approach mitigates the risk of structural reidentification through indirect relationships, which current methods do not consider.

NSA is also characterized by its ability to preserve graph connectivity while selectively anonymizing edges based on similarity thresholds. This combination of multihop protection and structure preservation allows for a better balance between privacy and analytical utility. Finally, NSA offers a flexible environment, adaptable to privacy and analytical needs, by allowing the selection of the anonymization depth (1-hop, 2-hop, or 3-hop), making it more suitable for practical applications in social network analysis.

5. Attack Scenarios and NSA's Defense Mechanisms

Social networks are subject to numerous reidentification attacks that an adversary could use to discover individuals within anonymized datasets. To prevent these, the Neighborhood-based Similarity Anonymization (NSA) technique utilizes multi-level anonymization, aiming at the direct and indirect associations to hide identifying patterns. The following are typical attack scenarios, and how the multilevel strategy utilized through the NSA supplies protection against each.

5.1. Neighborhood Attack

A neighborhood attack is a form that is founded on knowledge concerning the immediate neighborhood (1-hop relations) of an object to re-identify the object node from an anonymized network [11]. Such an attack is particularly strong if the first neighborhood of a node is an identifiable, unique pattern that is matchable to a known pattern with an adversary. Common anonymization techniques that anonymize direct relations only are often not secure against these attacks, as the indirect relations (e.g., 2-hop and 3-hop) could be identifiable as well.

The NSA thwarts attacks carried out at the neighborhood level by anonymizing at the 1-hop, 2-hop, and 3-hop neighborhood levels, thus modifying the relationships among immediate and distant neighbors. With the change of indirect relations, the individuality of each node structure at some levels of neighborhood is reduced, and thus neighborhood-based reidentification is greatly facilitated [8, 10].

5.2. Structural Attack (Subgraph Matching)

Structural attacks, or subgraph matching attacks, are based on the discovery of distinct structural patterns in a neighborhood of a node, frequently utilizing distinct subgraphs or clusters to match anonymous nodes with known structures [14]. In the attack, the adversaries utilize the distinct patterns to re-establish the users in anonymous datasets through matching identifiable subgraphs.

NSA defends against structural attacks through the anonymization of individual neighborhood levels, breaking the patterns at subgraphs that the adversaries utilize for matching. Through the alteration of 1-hop, 2-hop, and 3-hop neighborhoods, the uniqueness of a node's subgraph is reduced, resulting in the dilution of unique patterns and the defense against structural reidentification [9].

5.3. Recursive or Multihop Attack

Recursive or multihop attacks exploit relationships that extend beyond direct connections, searching 2-hop and 3-hop neighborhoods to identify individuals within anonymized networks [15]. These attacks seek to exploit the overall network structure, rendering single-level anonymization techniques highly susceptible, as these do not alter indirect relations. NSA's multi-level anonymization targets explicitly the multihop connections that are the backbone for defending against recursive attacks. The 2-hop and 3-hop anonymization at the edge hides patterns of indirect relations, so the adversaries cannot exploit extended neighborhood relations for reidentification [12, 16]. The multi-level design thus gives an efficient defense against attacks for recursive reidentification.

5.4. Edge-Based Attacks

Edge-based attacks focus on identifying specific links in the network that would reveal sensitive relations [17]. For social networks, some links are unique and could serve as identifiers for reidentification. Traditional edge-oriented anonymization approaches generally exclude or modify edges to prevent this risk; however, these do potentially interfere with meaningful properties of the network. NSA selectively modifies or removes edges based on similarity scores across neighborhood levels, effectively anonymizing sensitive edges while maintaining the overall network structure. By altering high-similarity edges across 1-hop, 2-hop, and 3-hop neighborhoods, NSA preserves network connectivity while protecting privacy against edge-based attacks.

5.5. Summary

The NSA approach provides robust protection against various reidentification attacks exploiting direct and indirect links in social networks. With independent anonymization for each neighborhood level, NSA effectively defends neighborhood, structural, recursive, and edge-based attacks, and balances the tradeoff between privacy and utility.

The multi-layer strategy ensures that sensitive relationships are protected, structural coherence is maintained, and a new state-of-the-art for privacy-aware anonymization is established for social networks.

6. Proposed Approach

6.1. Overview

The Neighborhood-Based Similarity Anonymization (NSA) framework outlines a multi-layered approach for anonymizing social networks by examining and transforming similarities between nodes within diverse network neighborhoods, with a particular focus on 1-hop, 2-hop, and 3-hop proximities. Traditional anonymization approaches often rely entirely on direct (1-hop) connections, which riskily overlook important indirect ties that an adversary could leverage for purposes of reidentification. The NSA framework addresses this deficiency by providing a more comprehensive view of each node's relational landscape, anonymizing relations based on both direct and extended ties.

6.2. Methodology

6.2.1. Data Preprocessing

Raw social text requires preprocessing to clean it up for analysis. Preprocessing for NSA involves normalizing the text content and removing unnecessary information to identify relevant linguistic attributes.

- Data loading:
  Raw data, that is, the relations and the tweet contents, are imported directly from the CSV file. The row is for a tweet pair and includes the source ID and the target ID for each tweet.
- Stop words Removal:
  In NLTK, stop words are imported to exclude common but unimportant words from the text. The deletion of stop words enables the retention of meaningful content per tweet.
- Data Cleaning and Normalization:
  - Removing URLs and Mentions
    URLs and user mentions (e.g., usernames) are excluded to reduce noise, as they do not contain useful content.
  - Special Characters and Case Normalization
    Non-alphanumeric characters are eliminated, and the entire text is normalized to lowercase to gain consistency, hence making the comparison accurate.
- Tokenization
  Each tweet is decomposed into individual lexical units (tokens) to facilitate independent analysis for each word.
- Lemmatization
  Words are simplified to their root forms through the application of lemmatization, thereby ensuring that various inflections of an identical term (e.g., running versus run) are consistently regarded, which improves uniformity in similarity assessments.

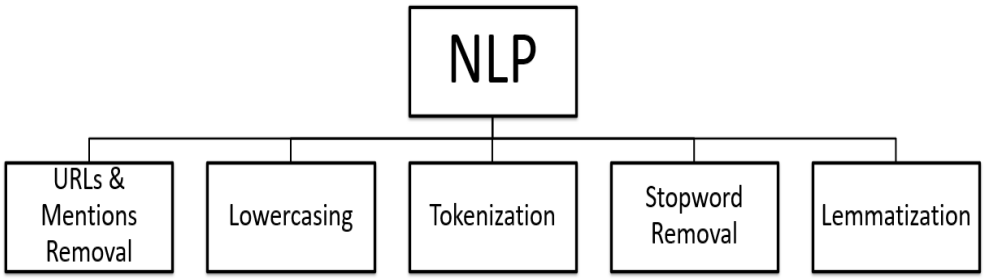

Fig. 1 Overview of NLP preprocessing techniques

6.2.2. Similarity Computation

Recursive similarity is computed between tweet pairs to quantify relations, and this is the basis for edge change.

- Cosine Similarity
  This cosine similarity $Sim_{cos}(a,b)$ It is then computed accordingly for two vectors A and B, namely the tweets' post-vectorization:

$$Sim_{cos}(a,b) = \frac{A.B}{\|A\|.\|B\|} \quad (1)$$

- TF-IDF Weighting
  TF-IDF (Term Frequency-Inverse Document Frequency) balances the weightage of individual words with the frequency of the words in the documents, and it assigns more weight to the rare words that are more qualitative.

- Multi-Level Neighborhood Similarities
  - 1-Hop Similarity Sim_{1-hop}
    When an edge joins nodes u and v, it quantifies direct Similarity among them.

- 2-Hop Similarity Sim_{2-hop}
  This measures the extent to which the nodes u and v are similar if they are connected through an intermediary node w:

$$Sim_{2-hop}(u,v) = \frac{Sim_{1-hop}(u,w) + Sim_{1-hop}(w,v)}{2} \quad (2)$$

- 3-Hop Similarity Sim_{3-hop}
  Determines the Similarity between two nodes when they are connected by two intermediaries:

$$Sim_{3-hop}(u,v) = \frac{Sim_{1-hop}(u,w) + Sim_{1-hop}(w,x) + Sim_{1-hop}(x,v)}{3} \quad (3)$$

Each Similarity is maintained in a separate column, allowing for the selection of which to keep confidential.

6.2.3. Threshold Determination

A threshold is established for each degree of Similarity based on the mean similarity value.

$$Mean_{sim_{k-hop}} = \frac{1}{|E|} \sum_{(u,v) \in E} Sim_{k-hop}(u,v) \quad (4)$$

6.2.4. Edge Deletion and Graph Anonymization

The NSA method eliminates, in a selective manner, edges that are too similar, using a predefined threshold. For each edge (u, v), if the 1-hop Similarity of the nodes is higher than the average Similarity at k-hop, then the edge is eliminated. The connectivity of the resulting graph is maintained to prevent the loss of the underlying structure for the benefit of anonymization, through an auxiliary copy of the resulting graph. The edge is eliminated only if, without it, the connectivity is not severed. The structural integrity of the network is not violated through this process.

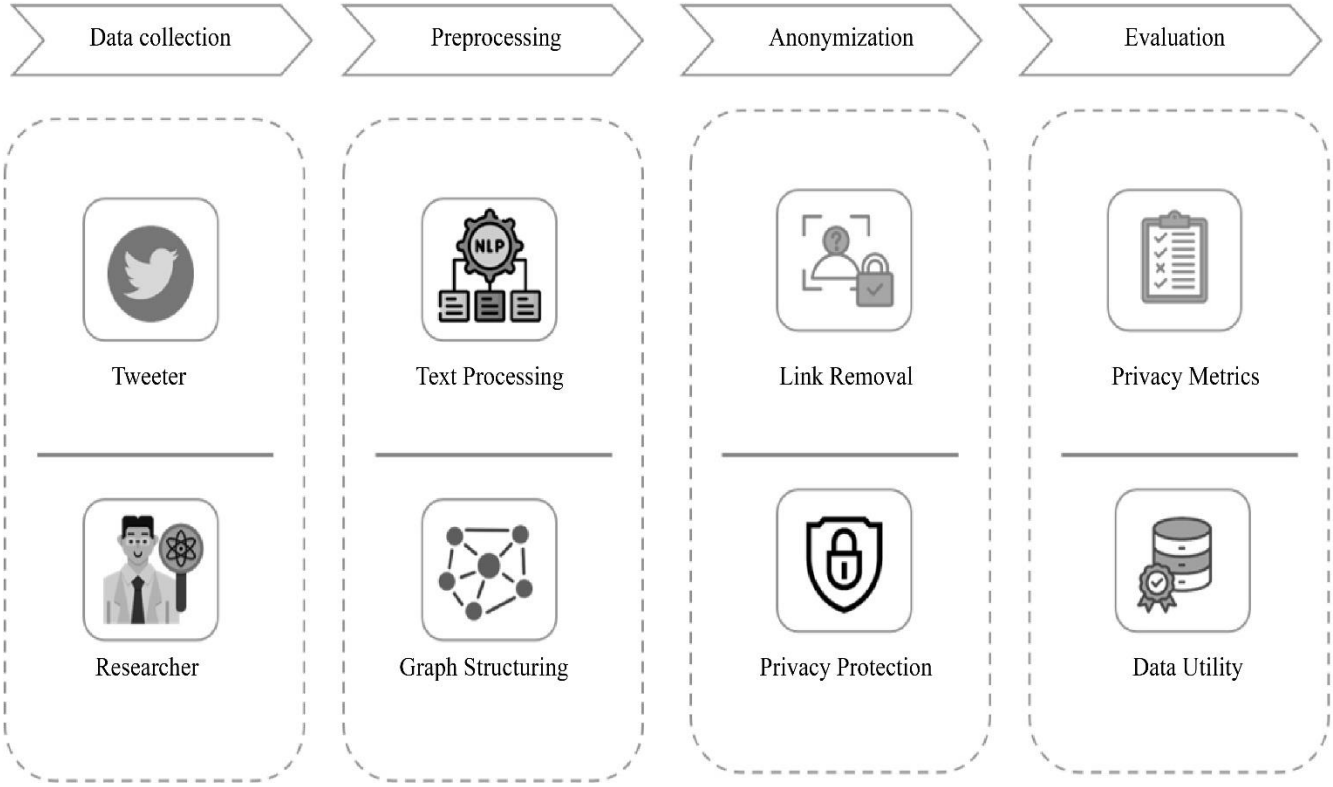

Fig. 2 Proposed data anonymization methodology

6.3. Algorithmic Complexity and Scalability

The Neighborhood-Based Similarity Anonymization (NSA) method relies on text preprocessing, similarity calculations, multihop analysis, and structural graph checks. Estimating its complexity demonstrates its ability to handle large social networks.

6.3.1. Text Preprocessing Complexity

Let N be the number of users and T the total number of tokens in tweets.

- Tokenization $O(T)$
- Stopword removal/normalization $O(T)$
- Lemmatization $O(T \,.\, c)$where c is the average cost of accessing the dictionary
- TF-IDF vectorization $O(N \,.\, V)$ where V is the vocabulary size

Overall preprocessing complexity

$$O(T + N \,.\, V) \quad (5)$$

This step scales linearly with dataset size, making it feasible for millions of tokens.

6.3.2. Similarity Computation Complexity

Let E be the number of edges, and the cosine similarity between all connected pairs

$$O(E \,.\, d) \tag{6}$$

Where d is the dimensionality of TF-IDF vectors.

Extending Similarity to multihop

- 1-hop similarity $O(E)$
- 2-hop similarity: $O(E \,.\, k)$ where k is average node degree
- 3-hop similarity: $O(E \,.\, k^2)$

So multihop similarity growth is polynomial in k, but still manageable for sparse networks (as Twitter tends to be).

6.3.3. Edge Removal and Connectivity Preservation

For each edge, NSA checks:

- Similarity threshold exceedance
- Connectivity preservation via temporary graph copy

Connectivity check using DFS or BFS costs:

$$O(N + E) \tag{7}$$

Since this is applied to each candidate edge E_p :

$$O\left(E_p\,(N + E)\right) \tag{8}$$

However, most graphs are sparse ($N \approx E$), making the practical cost much lower. Connected components are cached, which significantly reduces redundant checks.

6.3.4. Scalability Considerations

NSA is highly scalable because

- Graphs are sparse
- Multihop neighborhoods rarely explode in real-world social networks
- TF-IDF is optimized for sparse matrices
- All computations can be parallelized (vectorization, Similarity, hop scans)

In experiments, NSA processed 3,474 nodes and 2.6M edges efficiently on a standard machine, demonstrating its practicality.

7. Evaluation and Comparison

7.1. Dataset

In this work, a real Twitter dataset was used to mine the knowledge in it to study the problem and solution of social network anonymization. Twitter, with over 330 million monthly active users, provides a heterogeneous, large-scale, and dynamic ecosystem, making it a suitable environment for studying privacy risks and assessing anonymization techniques. The dataset in this study was acquired via personal communication with a researcher who had collected and organized the data for academic research. While the original collector could not be reached for formal attribution, the dataset is intact, structured, and research-ready, offering genuine relational and textual content for user interaction analysis without revealing PII.

The data is presented in the form of several CSV files, which contain user profiles, tweets, and connections. Profile fields, such as names, locations, and descriptions, are part of user profiles. Tweet files contain the text that users write, and each user appears a few times, depending on their level of activity. Relationship files encode social relations as pairs of user IDs that show friendship or follow relationships. This study did not require directionality; thus, undirected edges were generated from the friendship file to streamline the graph while preserving essential relational information.

Before the analysis, it was necessary to integrate the data from several folders. To get one representation of each user for the similarity calculation in the anonymization, all of the user's tweets were combined into one text entry. The final dataset is a well-organized and consistent collection of 3,474 accounts, 8,377,522 tweets, 34 user features, and 2,662,277 undirected edges. This comprehensive architecture provides a solid foundation for preprocessing, feature extraction, and the subsequent application of the Neighborhood-Based Similarity Anonymization (NSA) framework.

7.2. Edge Deletion

NSA's output is assessed through the following indicators:

Edge deletion ratio indicates structural changes, with a higher ratio indicating a higher percentage of anonymization.

$$ProportionRemoved = \frac{Initial\ edge\ number - Final\ edge\ number}{Initial\ edge\ number} \tag{9}$$

Average Path Length (APL) reflects network connectivity. A lower APL signifies better connectivity, whereas a significant increase implies reduced connectivity.

$$APL = \frac{1}{|V|(|V|-1)} \sum_{\substack{u \neq v \\ u,v \in V}} d(u,v) \tag{10}$$

Graph Connectivity Ensures that the graph remains a single connected component, preserving usability for analytical purposes.

7.3. Connectivity

The NSA method is also contrasted with two other alternative approaches: Whole-Graph Anonymization: It deletes edges at a constant rate without concerning itself with specific levels of neighborhoods.User Similarity-Based Anonymization [1]. It removes 25.6% of the edges while maintaining connectivity and without eliminating any nodes.

It is concerned with the direct user link and achieves protection for privacy without experiencing drastic connectivity loss.

The findings, encapsulated in Table 1, demonstrate the efficacy of NSA's 1-hop, 2-hop, and 3-hop anonymization strategies in relation to various alternative techniques.

Table 1. Comparative evaluation of NSA (1-Hop,2-Hop, and 3-Hop), whole-graph, and M.Ramdi et al.'s similarity-based anonymization

Anonymization Method	Edges Removed	Proportion of Edges Removed	Connectivity	Node Removal	APL (Original)	APL (Anonymized)
NSA- 1-Hop	114	11.41%	Preserved	0	3.1591	2.6805
NSA- 2-Hop	235	23.52%	Preserved	0	3.1591	2.3926
NSA- 3-Hop	296	29.63%	Preserved	0	3.1591	2.5543
Whole-Graph	215	21.50%	Reduced	0	3.1591	3.6275
Proposed Approach	215	25.6%	Preserved	0	3.1591	3.8095

8. Results

8.1. Proportion of Edges Removed

The 2-hop and 3-hop strategies proposed by the NSA reduced a larger percentage of the edges (23.52% and 29.63%, respectively) than the solution proposed in [1] (25.6%) and the whole-graph strategy (21.50%). Such flexibility enables the NSA to perform robust anonymization here, where it is necessary, and the 1-hop strategy to the NSA to prune edges (lower) at 11.41% is a positive if keeping more edges is the top priority. The algorithm proposed in [1] prunes 25.6% of the relations, focusing on direct relations. Although this method achieves a very high degree of anonymization, it deletes the indirect relations and the higher-level network dynamics observed with the multi-layered view of NSA.

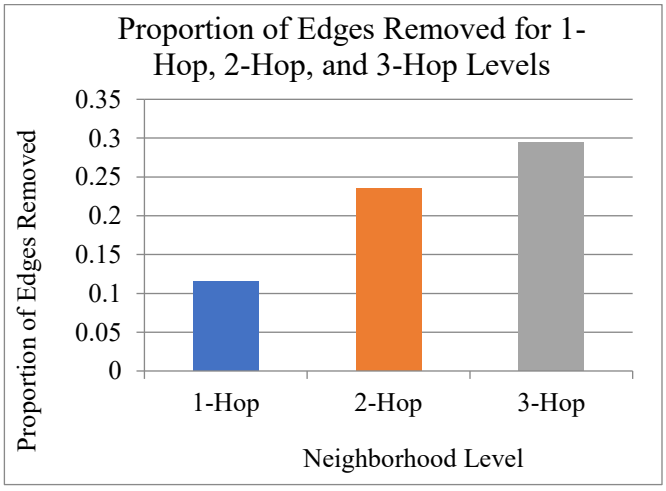

Fig. 1 Proportion of edges removed for 1-hop, 2-hop, and 3-hop levels

8.2. Average Path Length (APL)

NSA reduced the Average Path Length (APL) for all levels, with the greatest reduction seen for the two-hop NSA method (APL decreased from 3.1591 to 2.3926). The reduction demonstrates an increase in connectivity, suggesting that the NSA, in addition to securing privacy, can also increase the efficiency of the network. Whole-graph anonymization led to an increase in the APL from 3.3564 to 3.6275, indicating a reduction in connectivity. The methodology outlined in [1] maintained the Average Path Length (APL) at a moderate value, with a marginal increase from 3.3564 to 3.8095. Although it retains the connectivity of the graph, it cannot realize the extent of connectivity improvement that is achieved with the NSA 2-hop method.

8.3. Graph Connectivity

NSA's multi-level anonymization (particularly the 2-hop level) retains or enhances connectivity, verified with the lower APL. Such an outcome indicates the potential of the NSA to find a balance point between privacy and network usability.

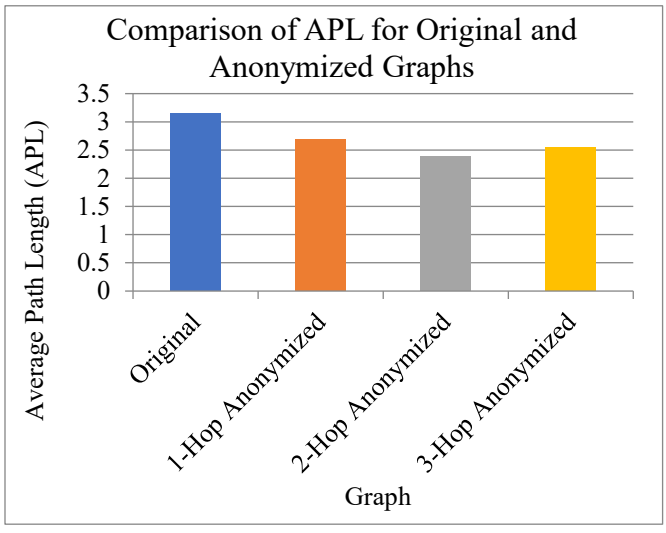

Fig. 2 Comparison of Average Path Length (APL) for original and anonymized graphs across 1-hop, 2-hop, and 3-hop Levels

The whole-graph method is also negatively impacting connectivity, characterized by the rise of APL. It is thus less suited to applications that require efficient network performance. The process presented for [1] retains

connectivity and circumcents the deletion of nodes, similar to the NSA strategy. Nevertheless, it lacks the connectivity improvements provided by the 2-hop strategy of NSA.

8.4. Utility and Flexibility of NSA

By offering 1-hop, 2-hop, and 3-hop levels of customization, the NSA supports an extent of privacy decisions, thus supporting dynamic anonymization strategies based on the desired balance between privacy and utility. Here, the 2-hop NSA method is notable for the fact that it not only strengthens connectivity but also efficiently anonymizes most of the edges. Though efficient, M. Ramdi et al.'s method is not as versatile as the multi-level anonymization of NSA, since it is mostly concerned with direct relations.

8.5. Mean Similarity

Mean Similarity illustrates the efficiency of our Neighborhood-Based Similarity Anonymization (NSA) technique for the scenario of graph anonymization. Prior to the anonymization process, the average Similarity at the 2-hop neighborhood level reaches its peak value of 0.14, indicating a perfect correlation between nodes that are indirectly connected (e.g., friends of friends), making the relations highly susceptible to inference attacks. After the anonymization process, a decrease in Similarity is observed at all neighborhood levels (1-hop, 2-hop, and 3-hop), indicating a clear change in the relations among the nodes. This approach is particularly noteworthy because it focuses on critical similarity levels, specifically the 2-hop neighborhood. Such a concentration significantly reduces the risk of reidentification, retaining the fundamental graph structure. The retention of global patterns, as evidenced by the robustness of the curve shapes, ensures that the resulting anonymized dataset is valuable for future analysis, such as community identification and network modeling. Compared to standard approaches, our NSA method strikes a balance between the best protection for sensitive information and data utility, making the anonymized graph both analytically useful and safe.

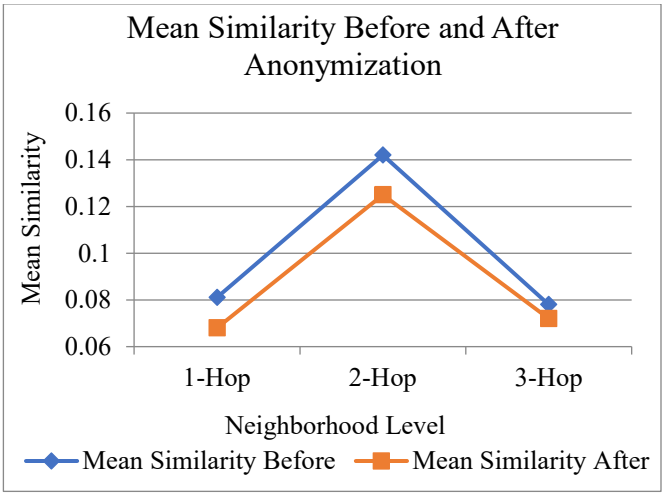

Fig. 3 Mean similarity before and after anonymization

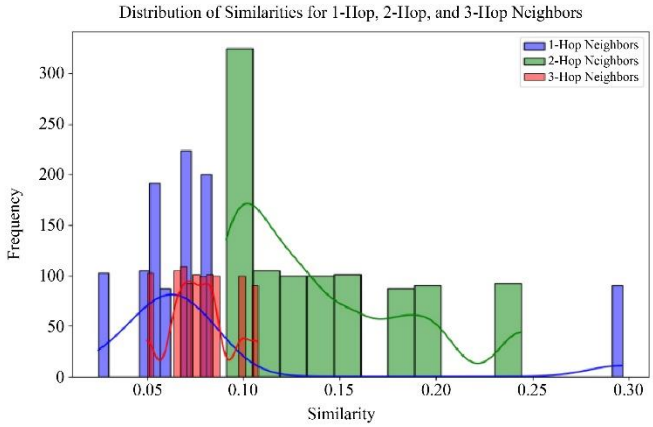

Fig. 4 Distribution of similarities for 1-hop, 2-hop, and 3-hop neighbors

9. Discussion

The Neighborhood-Based Similarity Anonymization (NSA) framework demonstrates notable strengths in striking a harmonious balance between privacy protection and data utility in the context of social network anonymization. Unlike traditional approaches, which often find themselves tilting toward privacy or connectivity, the multi-level anonymization framework offered by NSA allows for flexibility using several neighborhood tiers (3-hop, 2-hop, 1-hop), thus enabling accommodation with the variety of anonymization requirements. Such flexibility is particularly evident in the results regarding the elimination of an edge, Average Path Length (APL), and connectivity.

2-hop and 3-hop approaches for the NSA involve reducing the number of edges to a greater extent than the whole-graph strategy and similarity strategy proposed by M. Ramdi et al. The increased ratio of edge deletion enables a more effective anonymization process by deleting identifiable connections, particularly those with sensitive characteristics such as 2-hop and 3-hop connections. However, the 1-hop NSA strategy allows for a lesser number of edge deletions, making it an appropriate solution for those cases requiring higher connectivity and structural Similarity. The flexibility inherent in the edge selections reflects the flexibility of the NSA, demonstrating its applicability to the entire range of privacy settings, from those with minimal to those with extensive anonymization requirements.

Further, the Network Structure Anonymization (NSA) process reveals significant enhancements in Average Path Length (APL), especially with the adoption of the 2-hop strategy. The APL decreases from 3.1591 to 2.3926 for the 2-hop strategy of NSA, indicating an increase in network connectivity, with a generally shorter length, which signifies more feasible pathways for communication among the nodes. The findings presented here directly contradict those for whole-graph anonymization and the results published by M. Ramdi et al., which show either an increase in APL or a marginal reduction. The APL reduction observed with the

process through NSA not only indicates successful anonymization but also that the process has the potential to preserve, if not enhance, network operability for analysis tasks, indicating the best usage thereof under circumstances where efficiency in the network is paramount. In terms of connectivity, the NSA is more efficient than the whole-graph method, albeit at the cost of reduced connectivity, leading to inflation of APL and a decrease in node accessibility. The 2-hop NSA method, on the other hand, strikes a balance, maintaining network connectivity while ensuring robust positive anonymization. Such a balance is crucial for social network analysis, where maintaining network cohesion and usability after anonymization is often necessary for realistic and functional analysis. The method presented by M. Ramdi et al. is efficient in maintaining the direct relationship. However, it lacks the multi-level flexibility provided by the NSA, and thus, it does not offer overall protection from indirect relationship-based attacks.

The NSA's capacity to support degrees of anonymization to some extent, specifically in 1-hop, 2-hop, and 3-hop configurations, is indicative of its efficiency and adaptability. The inherent customizable functionality of NSA allows users to choose an anonymization degree that will best suit the desired balance between privacy concerns and the utility of the data. Such adaptability is particularly beneficial in contexts where datasets require varying levels of anonymization. Notably, the 2-hop method of NSA is successful because it can anonymize most of the edges with efficacy, also increasing the connectivity at the network level, which is not often accomplished with classical approaches. Although the method proposed by researcher M. Ramdi et al. is efficient, it is restrictive due to its focus on direct relations, rather than the intricate patterns of relations that are successfully tackled with the multi-level method of the NSA.

10. Conclusion

Neighborhood-Based Similarity Anonymization (NSA) is a big step forward in keeping social networks private. NSA offers a full and flexible solution that meets privacy needs while still being useful to the network by using multi-level analysis (1-hop, 2-hop, and 3-hop). In the past, people had to choose between privacy and utility. But with NSA, users can choose how much anonymity they want, making it a scalable system. The 2-hop level is interesting because it lets you hide a large number of edges while also making the network more connected. This is a good balance between privacy and network efficiency. NSA's flexibility enables the identification of both direct and indirect relationships within the network, overcoming the limitations of single-level approaches that ignore overall relational dynamics. This comprehensive view is essential for providing credible privacy protection against a wide range of attack patterns, including those exploiting multihop relationships. NSA's ability to decrease Similarity while maintaining or increasing connectivity offers practical applications in areas such as social network analytics, targeted advertising, or any scenario where network structure must be preserved despite anonymization.

However, limitations must be acknowledged: NSA may have scalability limitations on very large and dynamic networks, and non-selective NSA may introduce trade-offs in certain dense graph structures or sub-communities. Future work should explore optimization techniques to handle larger and more complex networks, evaluate NSA in various real-world use cases, and generalize NSA to other types of graphs (directed, weighted) to improve its generality and robustness.

In summary, NSA is a robust, flexible, and efficient solution to the long-standing problems of social network anonymization and could set new standards for privacy-preserving network analytics.

Appendix A: Glossary of Technical Terms

Node	A single user or entity represented in the social network graph
Edge	The connection or relationship (e.g., friendship or follow) between two nodes.
Hop / Neighborhood	The distance between two nodes in a network. A 1-hop neighbor is directly connected; a 2-hop neighbor is connected through one intermediary; and so on.
Graph Anonymization	The process of modifying a social network's structure to prevent user reidentification while retaining analytical utility
Similarity	A measure (often cosine similarity) indicating how alike two users are based on their tweet content or behavioral patterns.
Utility	The usefulness of anonymized data for further analysis or applications.
Privacy	The protection of individual identities against reidentification in anonymized data.

References

[1] R. Mariam et al., "An Innovative User Similarity-Based Privacy Preservation Approach," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 17, 2024. [Google Scholar] [Publisher Link]

[2] Latanya Sweeney, "k-Anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557-570, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ashwin Machanavajjhala et al., "L-Diversity: Privacy Beyond k-Anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3-es, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and L-Diversity," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, Turkey, pp. 106-115, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[5] Cynthia Dwork, "Differential Privacy," *Automata, Languages and Programming*, pp. 1-12, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[6] Michael Hay et al., "Anonymizing Social Networks," *Computer Science Faculty Publication Series, University of Massachusetts Amherst*, pp. 1-18, 2007. [Google Scholar]

[7] Sen Zhang, Weiwei Ni, and Nan Fu, "Differentially Private Graph Publishing with Degree Distribution Preservation," *Computers & Security*, vol. 106, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Bin Zhou, and Jian Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks," *2008 IEEE 24th International Conference on Data Engineering*, Cancun, Mexico, pp. 506-515, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[9] Kun Liu, and Evimaria Terzi, "Towards Identity Anonymization on Graphs," *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 93-106, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[10] Michael Hay et al., "Resisting Structural Re-Identification in Anonymized Social Networks," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 102-114, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[11] Arvind Narayanan, and Vitaly Shmatikov, "De-Anonymizing Social Networks," *2009 30th IEEE Symposium on Security and Privacy*, Oakland, CA, USA, pp. 173-187, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[12] Alina Campan, and Traian Marius Truta, "Data and Structural k-Anonymity in Social Networks," *International Workshop on Privacy, Security, and Trust in KDD*, pp. 33-54, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yang Li et al., "Private Graph Data Release: A Survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1-39, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Xiaowei Ying, and Xintao Wu, "Randomizing Social Networks: A Spectrum Preserving Approach," *Proceedings of the 8th SIAM International Conference on Data Mining (SDM)*, pp. 739-750, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[15] Chris Clifton, "Privacy-Preserving Data Mining," *Encyclopedia of Database Systems*, Springer, New York, pp. 2819-2821, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] Vibhor Rastogi et al., "Relationship Privacy: Output Perturbation for Queries with Joins," *PODS '09: Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 107-116, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sean Chester, and Gautam Srivastava, "Social Network Privacy for Attribute Disclosure Attacks," *2011 International Conference on Advances in Social Networks Analysis and Mining*, Kaohsiung, Taiwan, pp. 445-449, 2011. [CrossRef] [Google Scholar] [Publisher Link]