Journal of Theoretical and Applied Information Technology
15th September 2024. Vol.102. No. 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN INNOVATIVE USER SIMILARITY-BASED PRIVACY PRESERVATION APPROACH

MARIAM RAMDI[1], OUMAIMA LOUZAR[2], OUAFAE BAIDA[3], ABDELOUAHID LYHYAOUI[4]

[1,2,3,4] LTI Lab, ENSA of Tangier, Abdelmalek Essaâdi University, Tangier, 90000, Morocco

E-mail: [1]mariam.ramdi@etu.uae.ac.ma, [2]oumaima.louzar@etu.uae.ac.ma, [3]wbaida@uae.ac.ma, [4]lyhyaoui@gmail.com

ABSTRACT

Social networks are pivotal in various domains such as e-commerce, healthcare, and politics, providing valuable data for numerous applications. However, leveraging this data for tasks like recommendation systems and decision-making often encounters challenges related to user privacy. This paper proposes a novel approach to privacy preservation that centers on user similarity within social network graphs. Our method addresses the dual objectives of safeguarding user privacy and maintaining data utility. By prioritizing similarity among users, our approach effectively reduces information loss and enhances the accuracy of the results. This contribution is significant in advancing the responsible use of data in social network analyses, ensuring both privacy protection and high-quality information retrieval.
Keywords: *Anonymization, Similarity, Link removal, Social networks, NLP*

1. INTRODUCTION

With the advent of social networks, individuals worldwide have become interconnected, sharing information, and interacting with each other in unprecedented ways. These platforms have revolutionized communication, sharing experiences, and engaging with the world. However, this proliferation of personal data has also raised concerns about protecting privacy and securing users' sensitive information.

One of the primary challenges in this context is the anonymization of data in social networks. Anonymization aims to dissociate personally identifiable information from shared data, ensuring that individuals can no longer be directly identified. This is crucial for preserving user confidentiality and reducing the risks associated with disclosing their personal information.

Despite the critical importance of data anonymization, existing methodologies often fall short of maintaining the balance between data utility and user privacy. They may either overly generalize the data, leading to a loss of valuable information, or inadequately protect privacy, increasing the risk of re-identification.

To address these challenges, this article presents an innovative method for data anonymization in social networks based on user similarity. The primary purpose of this study is to develop and evaluate a robust anonymization technique that enhances privacy protection while preserving data utility. By leveraging user similarity, the proposed method aims to improve the effectiveness of anonymization, reduce the risk of re-identification, and ensure that anonymized data retains its value for analysis. Additionally, the method incorporates advanced techniques such as natural language processing (NLP) for tweet vector representation and similarity computation, adding a nuanced layer to the anonymization process. The efficacy of the proposed method is validated through comprehensive evaluation using real datasets from popular social networks.

2. RELATED WORK

Privacy in Online Social Networks (OSNs) is an emerging research domain that is still in the process of maturation. Most of the research in this field primarily adopts a computational perspective. This discussion will focus on several methods that have been proposed and remain relevant to OSN privacy. The prevalent paradigm in OSN privacy research conceives an OSN as a network consisting of nodes and links. As articulated in scholarly works such as [1] and [2], there are three key facets

where privacy concerns must be addressed to ensure comprehensive OSN privacy: node privacy, link privacy, and attribute privacy.

In surveying the landscape of anonymization techniques within the realm of social networks, it is imperative to contextualize our proposed methodology within the broader body of related work. An array of anonymization methods has been developed, each navigating the delicate balance between preserving data utility and safeguarding user privacy. Classical approaches, such as k-anonymity [3], and l-diversity [4], have laid the foundation for understanding the intricacies of privacy-preserving techniques. These methods, however, often fall short of capturing the nuances of social network data.

Various models have been developed to anonymize simple graphs, each aimed at mitigating specific types of structural attacks. Among these, the degree attack is particularly significant, as it allows an adversary to identify an individual based on their degree in the social network graph. To address this vulnerability, the k-degree anonymity model has been introduced, as highlighted in [5]. This model modifies the graph's degree sequence to ensure that each degree appears at least k times, thereby enhancing anonymity. The degree sequence, which consists of the set of node degrees in the graph, is altered to achieve this goal. Numerous techniques have been proposed and further developed based on this model to improve privacy protection.

One notable approach is the rapid k-degree anonymization method proposed by Lu, Song, and Bressan [6]. This algorithm employs a greedy strategy to add edges to the graph simultaneously, thus accelerating the anonymization process. However, it is important to note that while this method enhances the speed of achieving anonymity, it also increases the edit distance between the anonymized graph and the original graph.

Hartung et al. [7] introduced an improved version of their previous algorithm, leveraging dynamic programming techniques to anonymize the graph's degree sequence. This iteration also incorporates a data reduction rule to enhance the algorithm's speed. However, it is important to recognize that this improvement results in an increased runtime for the algorithm. In the context of k-anonymity, Bredereck et al. [8] proposed an alternative method that involves adding vertices and their corresponding edges to the graph. In this algorithm, the added vertices are duplicates of existing ones, sharing the same neighbors. Additionally, the study highlights the weak NP-hard nature of the block sequence problem, where the block sequence refers to a group of vertices with identical degrees. A problem is considered weakly NP-hard if it can be solved in polynomial time.

Ma et al. [9] introduced a two-step algorithm for achieving k-anonymity in social network graphs. The first step of the algorithm determines the final degree of each vertex, and the second step modifies the graph to ensure each vertex reaches the specified degree. This method exclusively adds edges without removing any, which helps maintain the graph's utility better than previous algorithms. However, this exclusive addition of edges also results in significant structural changes. To mitigate this, a more balanced approach that includes both adding and removing edges can be employed. In line with this, Casas Roma et al. [10] proposed a recent algorithm for k-anonymity in social network graphs that demonstrates improved graph utility compared to earlier methods. This algorithm involves two main steps: first, it uses a graph partitioning technique [11] to anonymize the degree sequence, and second, it employs a greedy method to carefully select edges for addition and removal. Despite the enhanced graph utility, the runtime of the second step is significant, as it requires scanning all candidate edges for each modification, leading to multiple scans and increased computational time.

In 2017, a novel approach to k-degree anonymity was introduced to preserve the structural integrity of social networks while safeguarding individual privacy [12]. This algorithm addresses the challenge posed by significant structural changes in the graph when edges are established between distant vertices. To mitigate this issue, the algorithm prioritizes connecting vertices that are closely situated. The algorithm begins by partitioning graph nodes into clusters, emphasizing nodes with high connectivity and minimal binding. Initially, it selects a set of nodes with the highest degrees to act as cluster representatives, subsequently incorporating directly connected nodes. Each node's immediate neighbors are systematically evaluated, and those exceeding a predefined edge threshold with the cluster are assimilated. Following clustering, new edges are exclusively introduced within each cluster. The size of each cluster is meticulously controlled to adhere to the specified k-value in the k-anonymity model.

In 2018, Sharma and Pathak introduced an expanded version of the k-anonymity model that utilizes clustering methods for anonymizing users

within social network graphs. This approach involves grouping nodes based on their connectivity patterns within the graph: nodes with similar degrees are clustered together, while nodes with different degrees are assigned to separate clusters. To enhance privacy and mitigate potential attacks on clusters lacking internal edges, additional edges are introduced among these nodes. Similarly, in the same year, Zheng et al. [13] proposed an enhanced iteration of the k-anonymity model that also integrates clustering techniques. This model aims to further enhance user privacy within social network graphs through strategic clustering strategies.

In 2019, Kiabod et al. proposed an algorithm designed to optimize the runtime of degree sequence anonymization by implementing a tree structure derived from the graph. This innovative approach ensures consistent performance, even as the k value in the k-degree anonymity model increases. The tree structure proves to be adaptable for anonymizing degree sequences across a range of k values. Similarly, [14] introduced a K-In&Out-Degree Anonymity Algorithm tailored specifically for large dynamic social networks, with a focus on preserving the community structure of the graph. This algorithm strategically organizes the degree sequence using dynamic grouping rules to minimize disruptions to the graph's overall structure. When integrating a new node, the algorithm assigns it to a suitable group only if the chosen group meets the privacy criteria post-integration.

In their study published in [15], Hazra and Setua proposed a novel three-fold degree anonymity approach tailored specifically for trust circles within service communication entities. This method focuses on safeguarding identity, location, and behavior privacy concerns simultaneously. Trust circles are formed based on the cumulative impact of direct interactions surrounding the entity, ensuring comprehensive privacy protection through a combination of direct and indirect trust mechanisms, along with participation in 2-degree anonymity protocols. Another notable contribution discussed in [16] was presented by Kosari, Sardar, Abdollah, Mousavi, and Radfar, who introduced an innovative method combining fuzzy clustering with the firefly algorithm. Initially, their approach utilizes a modified K-member variant of the Fuzzy c-means algorithm to construct well-balanced clusters, each containing a minimum of k members. Subsequently, the firefly algorithm refines these clusters further, aiming to enhance both graph anonymization and data privacy within the network context.

Autors in [17] focused on the balance between privacy and utility in online social network (OSN) data sharing, especially in the context of machine learning applications. It proposes two anonymization approaches to protect individual privacy while ensuring data utility for third-party use. The first approach uses Generative Adversarial Networks (GAN) to generate anonymized graphs that balance both objectives. The second approach employs Integrated Gradient (IG) to identify and modify graph structures, safeguarding group-level information while disrupting attackers' ability to infer individual data. The methods are evaluated on real-world datasets, demonstrating effective privacy preservation without compromising data utility for legitimate purposes.

In 2023, Medková et al. addressed the challenge of preserving privacy in social network datasets while sharing them with third parties. It proposes a hybrid algorithm (HAkAu) that improves upon existing k-automorphism anonymization techniques, which safeguard against structural attacks by modifying the network structure. The paper introduces a genetic algorithm to handle the NP-hard problem of finding isomorphic graph extensions, and uses the GraMi algorithm for identifying frequent subgraphs. This method minimizes information loss and enhances computational efficiency, maintaining a balance between privacy and data utility. The proposed algorithm is tested on real-world social networks and compared against previous methods using the SecGraph tool [18].

Dehaki et al. [19] introduced a novel method for improving graph anonymization to protect user privacy in social network data. The proposed approach, CKH4KDA, combines the Chaotic Krill Herd (CKH) optimization algorithm with number factorization to efficiently select and delete graph edges, minimizing processing costs and maintaining the graph's structural integrity. The method significantly reduces execution time by scanning the graph only once and carefully selecting edges with minimal impact on the graph's utility. This algorithm outperforms existing techniques in both speed and utility preservation, as demonstrated through evaluations on multiple real-world datasets. The research highlights the balance between privacy protection and data utility, addressing the challenges of large-scale social network anonymization.

Authors in [20] tackled the challenge of publishing histograms of common neighbors in

social networks while ensuring edge-differential privacy. Common neighbor counts, which reveal key social network structures, are highly sensitive, making privacy preservation difficult. The authors propose a multi-stage approach that segments the histogram into parts with lower sensitivities, using the long-tail distribution of common neighbors to apply differential privacy more efficiently. This method balances privacy and utility by calibrating noise to each segment's sensitivity. Their experiments show improved utility for tasks like community detection, offering a novel solution for privacy-preserving data publication in social networks.

Recent advancements have witnessed the integration of machine learning and natural language processing (NLP) techniques into anonymization processes. Embedding-based models and clustering algorithms have shown promise in maintaining the structural integrity of graphs while obfuscating individual identities. Nonetheless, challenges persist, particularly in scenarios where preserving the inherent similarity between users is crucial for downstream analyses.

Our research builds upon these foundations, introducing a methodology grounded in user similarity. Leveraging NLP for tweet vector representation, our approach seeks to overcome the limitations of traditional methods. The emphasis on similarity as a guiding principle positions our methodology as a nuanced evolution within the landscape of social network anonymization. Comparative analyses with existing techniques serve to highlight the unique advantages and challenges associated with our proposed approach.

3. PROBLEM STATEMENT

In the context of this research, a graph-based modeling approach was adopted to investigate and analyze a social network. This methodology involved representing the social network as a graph G = (V, E), where individuals or entities are represented as nodes V, and the connections between them are illustrated as edges or links E. Notably, in this representation, the edges symbolize friendships or associations between users in the network. The graph employed in this context is undirected, signifying that nodes and edges carry no inherent weight. This graph model allows for the visualization of relationships within the social network, with each link denoting a friendship connection between the corresponding nodes.

The primary emphasis of this paper centers on the graph's structural characteristics, particularly focusing on the node degrees within the network. This graphical framework facilitated the application of well-established principles derived from graph theory and network analysis to conduct a comprehensive examination of the network's structural and dynamic attributes. Through the utilization of this graph-based model, various aspects were explored, including patterns of information dissemination, the identification of pivotal nodes, and the delineation of distinctive community structures within the social network.

The increasing complexity and volume of data in social networks have surpassed the capabilities of traditional privacy-preserving methods. Existing techniques often struggle to protect user privacy effectively, especially in dynamic environments characterized by intricate and evolving interactions. The literature indicates that many conventional approaches are inadequate for addressing the nuanced relationships and complex structures present in large-scale social networks.

This research addresses these limitations by adopting a graph-based model that incorporates similarity as the weight of edges, providing a more detailed and adaptable privacy-preserving mechanism. This approach offers a refined method for protecting user data while effectively analyzing network structures. The significance of this work lies in its ability to enhance both theoretical understanding and practical applications, offering more robust privacy protection strategies that are well-suited to the evolving landscape of social networks.

By incorporating similarity as the weight of the graph edges, the model captures the strength or intensity of the relations.

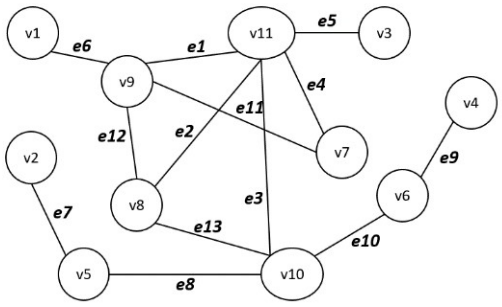

Figure 1: Social network representation as an undirected unweighted graph

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

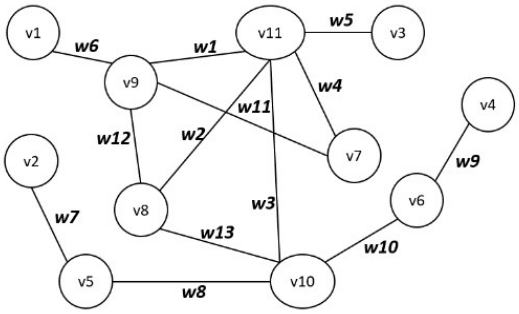

Figure2: Social network representation as an undirected weighted graph.

4. WHY SIMILARITY

Using tweet similarity provides several specific advantages in addressing privacy issues in social networks compared to traditional schemes (Table 1).

Using tweet similarity to calculate the weight between two nodes that are friends provides a sophisticated and effective approach to data anonymization in social networks. This method ensures contextual relevance, enhanced privacy protection, and better data utility, addressing many of the limitations of traditional anonymization schemes.

Table 1: Comparison of tweet similarity-based anonymization method vs. Traditional anonymization schemes

Aspect	Tweet Similarity-Based Method	Traditional Anonymization Schemes
Preprocessing	Utilizes NLP techniques for cleaning and normalizing tweet data.	Often involves basic data cleaning and standardization.
Similarity Calculation	Uses cosine similarity to measure the similarity between users' tweets based on their content.	May use generic measures of similarity or none.
Cut-off Determination	Uses the mean of similarity scores to determine a dynamic cut-off threshold.	Typically relies on fixed thresholds or less adaptive criteria.
Edge Deletion and Connectivity	Deletes edges with weight above the cut-off while ensuring the social network graph remains connected.	May remove or generalize data without considering network connectivity.
Enhanced Privacy Protection	Groups users with similar tweet content, making re-identification more difficult.	Generalizes data, which can be reverse-engineered more easily.
Data Utility Preservation	Maintains meaningful relationships and patterns in the data, preserving its value for analysis.	Often leads to significant data loss, reducing utility for analysis.
Contextual Relevance	Considers the actual content shared by users, leading to more accurate anonymization.	Lacks contextual relevance, treating all data points equally.
Scalability and Adaptability	Adaptable to dynamic data and scalable for large datasets; processes real-time tweet data efficiently.	Requires frequent re-evaluation and adjustment, making scalability challenging.
Reduction of Over-Generalization	Provides a finer balance by generalizing only as necessary.	Often over-generalizes to ensure privacy, reducing data usefulness.
Advanced Techniques Integration	Easily integrates with NLP and machine learning for enhanced precision.	May not readily accommodate advanced techniques.

5. PROPOSED APPROACH

Graph modification techniques involve the deliberate alteration of the social graph's structure by incorporating strategic adjustments, such as the addition and/or removal of nodes and edges [21]. Our approach, situated within this paradigm, introduces modifications to the graph structure with the explicit goal of preserving individual privacy. By strategically manipulating the connections and relationships represented in the social graph, our method aims to mitigate potential privacy vulnerabilities while maintaining the utility and integrity of the network.

The proposed method aims to address the critical challenges associated with privacy preservation in social network data while maximizing the utility of the information contained within. Focused on creating a nuanced and effective approach, the method strives to mitigate re-identification risks inherent in the vast and diverse datasets generated by social media platforms.

Leveraging advanced techniques in natural language processing and graph theory, the method seeks to navigate the complexities of data anonymization. By incorporating sophisticated similarity measures, statistical analyses, and thoughtful threshold determinations, the method aims to strike a delicate balance. It endeavors to ensure that the resultant anonymized data not only shields individual identities effectively but also retains the essential structure and meaningful patterns of the social network. In doing so, the proposed method aspires to provide a robust and adaptable solution that aligns with contemporary privacy standards in the ever-evolving landscape of online interactions.

This algorithm seamlessly integrates preprocessing techniques. Initial data cleaning ensures a standardized input, and similarity calculations facilitate the identification of significant relationships.

In the proposed approach, a novel strategy was implemented by substituting traditional friendship relationships with a measure of similarity between users. This departure from conventional social network modeling aimed to provide a more nuanced understanding of the connections within the network. By replacing binary friendship labels with quantitative similarity scores, the representation of relationships transitioned from a categorical to a more continuous and informative model. This paradigm shift allowed for capturing the inherent complexities and gradations in user interactions. The utilization of similarity scores as a relational metric offers a richer perspective, enabling the discernment of subtle variations in the strength and nature of connections between users. This innovative approach aligns with the evolving landscape of social network analysis, emphasizing a data-driven and nuanced representation of relationships for a more comprehensive exploration of user dynamics within the network.

The determination of a natural cutoff in the similarity distribution, based on statistical metrics, guides the removal of edges. This process prioritizes the preservation of graph connectivity while eliminating more relevant links. The algorithm culminates in a visually accessible final graph, demonstrating the efficacy of the applied techniques in balancing data utility and privacy preservation.

6. ALGORITHM

6.1 Preprocessing

The first step in our algorithm involves preprocessing the tweet data using natural language processing (NLP) techniques. This preprocessing stage is crucial for transforming raw tweet text into a clean and standardized format suitable for analysis. It includes several sub-steps

- URLs and mentions removal are advocated as these elements introduce extraneous noise and offer minimal contribution to textual analysis.
- Subsequently, a conversion to lowercase is proposed to normalize the text, ensuring uniformity in word comparison.
- Apply tokenization to break the tweet into discrete units or "tokens" for individual processing.
- Remove common words, referred to as "stop words," are subsequently eliminated to diminish semantic noise.
- To further streamline the text, lemmatization is employed, reducing variations to a base form, and simplifying the vocabulary dimensions.

Using tweet similarity provides several specific advantages in addressing privacy issues in social networks compared to traditional schemes.

6.2 Similarity calculation

After preprocessing, the next step is to calculate the similarity between users based on their tweets. In this research, various similarity metrics were systematically employed to scrutinize the connections among tweets within the social network. The metrics encompassed:

- Cosine similarity, which gauges the cosine of the angle between two vectors.

$$\text{cosine\_similarity}(a,b) = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}.\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

- Jaccard similarity, evaluating the intersection relative to the union of sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (2)$$

- Euclidean distance, determining the straight-line distance between points

$$d(p,q) = \sqrt{(p_1 - q)^2 + (p_2 - q_2)^2} \quad \text{P} \quad (3)$$

- TF-IDF Cosine Similarity, which considers term frequency and inverse document frequency to weigh term importance.

Figure 3 compares different similarity measures among tweets, namely TF-IDF Cosine Similarity, Jaccard Similarity, standard Cosine Similarity, and Euclidean Distance. Each measure is plotted against an index representing pairs of tweets.

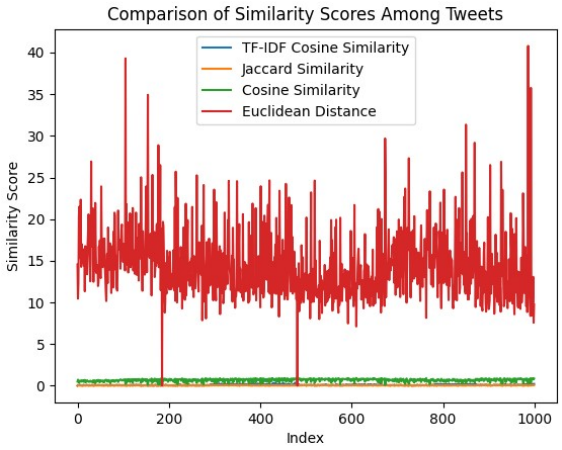

Figure 3: Comparison of similarity scores among tweets

TF-IDF Cosine Similarity scores are mostly close to zero, with slight variations among the tweet pairs. This indicates that the tweets have a high degree of sparsity and few common terms, which is typical in short texts like tweets. The TF-IDF weighting effectively reduces the impact of common but less informative words, focusing more on unique terms. This makes it a robust measure for capturing the relevant similarities between tweets by emphasizing the distinguishing features of the text.

The Jaccard Similarity scores are extremely low and nearly constant across the tweet pairs. This suggests that there is very little overlap in the sets of terms present in different tweets. Jaccard Similarity, which measures the intersection over the union of term sets, does not capture meaningful similarity in this context. This is likely due to the brief and varied nature of tweet content, where the overlap of terms between any two tweets is minimal, leading to uniformly low similarity scores.

Cosine Similarity scores, like TF-IDF Cosine Similarity, are also close to zero with minimal variations. Similar to TF-IDF Cosine Similarity, standard Cosine Similarity indicates low similarity between tweets. However, without the TF-IDF weighting, it might not be as effective in highlighting the more informative terms. Standard Cosine Similarity measures the angle between the vector representations of the tweets, but it treats all terms equally, which can dilute the impact of unique terms in the text.

Euclidean Distance shows a wide range of values, with scores varying significantly across the tweet pairs. The high variability in Euclidean Distance suggests it is heavily influenced by the magnitude of the vectors. This makes it less suitable for measuring tweet similarity, as it is affected by the length and frequency of terms, leading to less meaningful similarity scores for text data. Unlike cosine similarity, Euclidean Distance does not normalize for vector length, resulting in large distances that do not necessarily correlate with the actual content similarity between tweets.

Figure 4 highlights the effectiveness of cosine similarity measures, particularly TF-IDF Cosine Similarity, in providing stable and interpretable similarity scores for tweets. These measures are more robust and contextually relevant for text data compared to Jaccard Similarity and Euclidean Distance. This supports the choice of cosine similarity for calculating the weight between nodes in our algorithm, ensuring both privacy protection and data utility in the anonymization process.

Algorithm 1: Calculation of node similarity

```
Initialize
While (! EOF) do
    For (every tweet) do
        // Vectorization using a fictional NLP
        model
        Tweet_vectors = vectorize_tweets(tweet)

        // Cosine similarity calculation
        Similarity_matrix =
        cosine_similarity(tweet_Vector)
    End

    End
```

6.3 Threshold determination

From this distribution (Fig.4), the identification of a natural cutoff point, a threshold beyond which similarities are deemed significantly high, becomes pivotal. This determination involves a careful observation of the distribution's shape, pinpointing an area where the frequency of similarities experiences a sharp decline. According to the histogram's insights, the optimal threshold is established at 0.78. This thresholding process

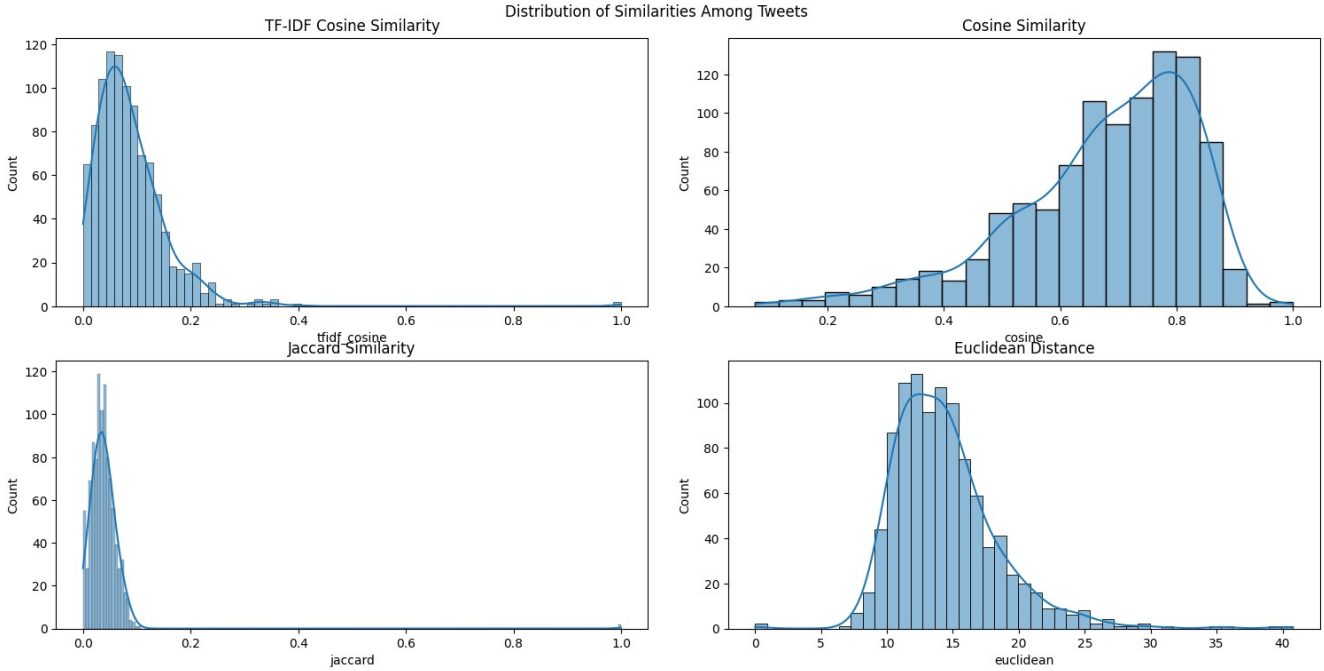

Figure 4: Comparison of similarity scores among tweets

constitutes a critical step in the methodology, delineating relevant similarities and guiding the selective removal of edges within the graph. The thus-determined threshold value serves as a crucial parameter to ensure effective anonymization while preserving the structural connectivity of the social network.

6.4 Edge deletion

In the algorithmic refinement process, the primary objective is to selectively eliminate strong connections within the graph based on a specified threshold. This strategic removal of edges surpassing the determined threshold is geared toward enhancing the discernibility of the network by eliminating relationships that may be considered robust but less informative.

The thresholding mechanism ensures that only connections exhibiting a similarity below the specified threshold are retained, contributing to a more refined and nuanced representation of the social network. Importantly, the algorithm incorporates a connectivity preservation strategy to ensure that the elimination of strong connections does not inadvertently disconnect the graph. This deliberate approach seeks to prioritize the preservation of meaningful and distinctive relationships while maintaining the overall connectivity of the network.

The result is a more focused and interpretable graph that facilitates a clearer andmore insightful analysis of the underlying social structure.

Algorithm 2: Edge removal

```
Initialize
While (! Edge.lenght()) do
    For (every edge) do
        // Identify edges with similarity values exceeding the threshold
         = get_strong_connections(graph, similarity_threshold)

        // Extract nodes involved in strong connections
        nodes_in_strong_connections = get_nodes_in_connections(strong_connections)

        // Remove edges with similarity values exceeding the threshold
        refined_graph = remove_strong_connections(graph, strong_connections)


        // Ensure that the edge suppression does not disconnect the graph
        preserve_connectivity(refined_graph, nodes_in_strong_connections)
    End
    End
```

7. DATASET

In this study, the wealth of information embedded within the Twitter database was harnessed to delve into the nuances of data anonymization. Twitter, a social media platform with over 330 million monthly active users, offers a diverse and dynamic dataset. By tapping into this extensive resource, the challenges and solutions associated with preserving user privacy while extracting valuable insights from the data were explored.

The Twitter database, rich in textual content, enabled navigation of the complexities of managing sensitive information. Simultaneously, cutting-edge anonymization techniques were employed to ensure the protection of individual identities. The analysis of the Twitter database serves as a real-world case study, highlighting the significance of data anonymization in the realm of social media and its implications for user privacy and data research.

The data under consideration is structured in a CSV format with columns denoted as "source_id", "target_id", "tweet_source_id", and "tweet_target_id". Each row in this dataset encapsulates the attributes of a pair of friends within the Twitter social network, thereby representing the interconnected tweets of these two individuals. Specifically, "source_id" corresponds to the identifier of the source node, indicating one friend, while "target_id" represents the identifier of the friend node, denoting the second individual. The columns "tweet_source" and "tweet_target" contain the respective tweets of the source and target nodes, providing textual insights into the online interactions and content shared by these connected friends. This data structure aligns with the context of social network analysis, enabling a granular examination of the relationships and shared content between pairs of friends on the Twitter platform.

This study underscores the delicate balance between extracting meaningful information and safeguarding user identities. It emphasizes the crucial role that advanced anonymization techniques play in maintaining this equilibrium, crucial for both data utility and privacy in the context of social media platforms.

Table 2: Dataset description

Dataset	Tweeter
Accounts	3474
Tweets	8,377,522
Attribute number	34
Edges	2 662 277

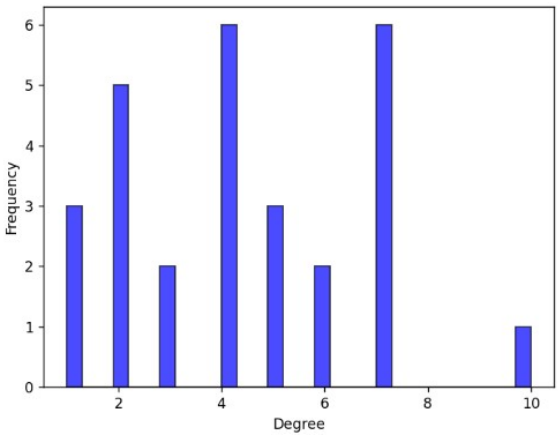

Figure 5: Degree distribution of the dataset

8. EVALUATION

In the evaluation of the proposed method, a comprehensive set of metrics and analyses was employed, encompassing link removal, graph connectivity, node removal, and average path length (APL). Link removal served as a crucial aspect of the assessment, focusing on the impact of eliminating specific connections or friendships within the social network graph. This allowed for gauging the resilience of the network structure to the removal of individual links, providing insights into the robustness and stability of the relationships [22].

Furthermore, graph connectivity analysis played a pivotal role in evaluating the overall cohesion and connectedness of the social network. Assessing the connectivity patterns after implementing the methodology provided a nuanced understanding of how well the network maintained its structural integrity and the extent to which the approach influenced the overall connectivity dynamics.

Additionally, the evaluation included node removal, wherein individual users or entities were systematically removed from the network. This aspect allowed us to explore the repercussions on network cohesion and information flow when specific nodes were eliminated. Together, these evaluation criteria provided a holistic perspective on the efficacy and implications of our proposed methodology, contributing to a comprehensive assessment of its impact on the social network structure and relationships.

The average path length (APL) of the original graph is 3.3564, while the APL of the anonymized graph is 3.8095. This modest increase of approximately 0.453 reflects the effectiveness of

our anonymization method in preserving the network's core connectivity and reachability. Despite the slight rise, the anonymized graph maintains a low APL, ensuring efficient information dissemination and social interactions. This demonstrates that our approach successfully balances privacy protection with network utility, retaining the essential small-world properties crucial for real-world applications.

The proposed similarity-based anonymization method has been rigorously evaluated, demonstrating a commendable balance between privacy preservation and utility in the context of a social network graph. Applying natural language processing (NLP) to preprocess tweets and extracting similarities between them, the algorithm successfully determined an optimal Similarity Threshold. The removal of edges with similarity values exceeding this threshold, while ensuring the preservation of graph connectivity, resulted in an anonymized graph with a 25.6% suppression rate. Importantly, the retention of the initial number of nodes attests to the method's efficacy in achieving anonymization without significant loss of network nodes.

This method's performance is further highlighted by its ability to maintain graph connectivity, crucial for preserving the overall structure of the social network. The moderate suppression rate of 25.6% indicates a nuanced approach, striking a balance between privacy enhancement and the preservation of meaningful relationships. The success of the algorithm in achieving its goals is evident not only in the suppression rate but also in the connectedness of the anonymized graph, reinforcing its utility for subsequent analyses.

The evaluation also considered the potential re-identification risks and privacy metrics, confirming the algorithm's effectiveness in safeguarding sensitive information. This method enables the extraction of valuable insights from the anonymized graph while minimizing the risk of user re-identification. Overall, the proposed similarity-based method presents a robust and adaptable solution for social network anonymization, making it a valuable tool for scenarios where data privacy and utility are of paramount importance.

Table 3: Data utility

Utility	Link Removal	Connectivity	Node removal	Average path length (APL)	
Proposed approach	25.6%	Preserved	No removed nodes	Original graph	Anonymized graph
				3.3564	3.8095

9. DISCUSSION

The incorporation of similarity as a key element is pivotal in addressing privacy concerns and enhancing user security. The algorithm's emphasis on detecting and eliminating links with strong similarity aligns with the overarching objective of mitigating the potential disclosure of sensitive information about users. This strategy recognizes the inherent risk associated with closely connected nodes, as such connections can inadvertently reveal significant relationships or affiliations.

Anonymizing data is fundamentally about finding a delicate balance between preserving individual identities and ensuring the data remains useful and informative. The deliberate removal of links with strong similarity becomes a strategic maneuver to minimize the risk of user re-identification and protect against the inadvertent disclosure of private details. By implementing this approach, the algorithm contributes to a more secure dataset, fostering user confidence in the protection of their personal information.

It is crucial to underscore that the integration of similarity-based anonymization in the algorithm underscores a commitment to both privacy and utility. The approach seeks to uphold the structural integrity of the graph while making it more challenging for adversaries to exploit relationships for user identification. This nuanced balance allows for continued data utilization for analytical purposes, ensuring that the anonymized graph remains a valuable resource for extracting meaningful insights.

However, the deployment of similarity-based anonymization methods necessitates careful consideration. Striking the right balance between privacy preservation and data utility require an understanding of the specific context and potential implications of re-identification risks. Continuous evaluation and refinement of such anonymization strategies are imperative to stay abreast of evolving privacy standards and ensure that the algorithm aligns with the ever-changing data privacy landscape.

10. CONCLUSION

The proposed algorithm has proven to be a significant advancement in the field of graph anonymization, offering a novel approach that preserves both the connectivity and integrity of the graph. Unlike traditional methods that often lead to substantial edge deletions and a loss of structural information, this algorithm strategically identifies and removes edges while maintaining the overall topology. This balance between anonymity and utility is critical, particularly in applications where the preservation of relationships and network structure is essential, such as social network analysis and relational data modeling.

The ability to retain all nodes within the graph further enhances the practical value of the algorithm, ensuring that the anonymized graph continues to reflect the full scope of the network's interactions. This is particularly significant in domains where metrics such as node centrality, clustering, and community detection play a pivotal role in understanding the network. By safeguarding the presence of all nodes, the algorithm not only achieves anonymization but also preserves the fundamental properties that make the graph a meaningful and representative model of the underlying data.

The experimental results validate the effectiveness of this approach, highlighting its capability to achieve a favorable balance between privacy preservation and data utility. This makes the algorithm a robust solution for privacy-conscious data handling, particularly in real-world applications where both anonymity and the structural integrity of the graph are paramount.

Beyond its immediate applications, this research lays the groundwork for future exploration in the area of node attribute anonymization. Future efforts will focus on anonymizing node attributes while maintaining strong similarities between the original and anonymized graphs. This extension will address the increasing need to protect not only the structural aspects of networks but also the sensitive information linked to individual nodes. Additionally, further optimization of the algorithm for larger, more complex datasets will be a crucial area of exploration, ensuring scalability and efficiency in diverse real-world scenarios.

In conclusion, this work presents a valuable contribution to the field of data anonymization in social networks, offering a solution that strikes a necessary balance between privacy and utility. As data privacy continues to be a critical concern, the proposed algorithm provides a practical and effective approach for preserving anonymity while retaining the essential characteristics of the network, paving the way for more secure and insightful analyses of social and relational data.

REFERENCES:

[1] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," IEEE Trans. Netw. Sci. Eng., to be published.

[2] Z. Cai and Z. He, "Trading private range counting over big iot data," in Proc. 39th IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS), Jul. 2019.

[3] L. Sweeney, "k-anonymity: a model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, pp. 557-570, 2002.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, pp. 24-24, 2006.

[5] Feder, T., Nabar, S. U., & Terzi, E. (2008). "Anonymizing Graphs,". CoRR, abs/0810.5,1–15. Retrieved from http://arxiv.org/abs/0810.5578v1.

[6] Lu, X., Song, Y., & Bressan, S. "Fast identity anonymization on graphs," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 7446, pp. 281–295, 2012.

[7] Hartung, S., Hoffmann, C., Nichterlein, A. A., Hoffman, C., Nichterlein, A. A., Hoffmann, C., & Nichterlein, A. A. "Improved upper and lower bound heuristics for degree anonymization in social networks," Lecture Notes in Computer Science, pp. 1–20, 2014.

[8] Bredereck, R., Froese, V., Hartung, S., Nichterlein, A., Niedermeier, R., & Talmon, N. "The complexity of degree anonymization by vertex addition," Theoretical Computer Science, vol. 607, pp. 16–34, 2015.

[9] Ma, T., Zhang, Y., Cao, J., Shen, J., Tang, M., Tian, Y., Al-Rodhaan, M. "KDVEM: A k-degree anonymity with vertex and edge modification algorithm," Computing, vol. 97(12), pp. 1165–1184, 2015.

[10] Casas-Roma, J., Herrera-Joancomartí, J., and Torra, V. "k-Degree anonymity and edge selection: Improving data utility in large networks," Knowledge and Information Systems, vol. 50(2), pp. 447–474, 2017.

[11] Hansen, S. L., & Mukherjee, S. "A polynomial algorithm for optimal univariate microaggregation,". IEEE Transactions on Knowledge and Data Engineering, vol. 15(4), pp. 1043–1044, 2003.
[12] Macwan, K. R., & Patel, S. J. "k-Degree anonymity model for social network data publishing," Advances in Electrical and Computer Engineering, vol. 17(4), pp. 117–124, 2017.
[13] Zheng, L., Yue, H., Li, Z., Pan, X., Wu, M., & Yang, F. "k-Anonymity Location Privacy Algorithm Based on Clustering," IEEE Access, vol. 6, pp. 1–1, 2018.
[14] Zhang, X. "Large-Scale Dynamic Social Network Directed Graph K-In & OutDegree Anonymity Algorithm for Protecting Community Structure," IEEE Access, vol. 7(1), pp. 1743–1768, 2019.
[15] Hazra, S., & Setua, S. K. "Privacy Preservation Using 2-Degree anonymity with trust circle in ubiquitous network for service communications,". IEEE Access, vol. 8, pp. 29965–29986, 2020.
[16] Kosari, R., Sardar, S., Abdollah, S., Mousavi, A., & Radfar, R. "Combined fuzzy clustering and firefly algorithm for privacy preserving in social networks," Expert Systems With Applications, 141, 2020.
[17] Gao, T., & Li, F. "Machine Learning-based Online Social Network Privacy Preservation," Asia Conference on Computer and Communications Security, pp. 467 - 478, 2022.
[18] Medková, J.,Hynek, J. "HAkAu: hybrid algorithm for effective k-automorphism anonymization of social networks," Soc. Netw. Anal. Min. 13, 2023.
[19] Toroghi, A. D., & Hamidzadeh, J. "Protecting the privacy of social network data using graph correction," Knowledge and Information Systems. pp. 1-33, 2024.
[20] Lv, C., Xiao, X., Zhang, L., & Yu, T. "Publishing Common Neighbors Histograms of Social Networks under Edge Differential Privacy," In Proceedings of the 19th ACM Asia Conference on Computer and Communications Security, pp. 1099-1113, 2024.
[21] R. Mariam, B. Ouafae, L. Oumaima, and L. Abdelouahid, "An Overview About Privacy Protection of Facebook Social Network Users Data," Advances in Intelligent Systems and Computing, vol. 1418, Springer, pp. 1132–1145, 2019.
[22] B. Ouafae, R. Mariam, L. Oumaima, and L. Abdelouahid, "Data Anonymization in Social Networks State of the Art, Exposure of Shortcomings and Discussion of New Innovations," The International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), 2020.