Comparison and Analysis of Distortion Techniques in Terms of Mitigating the Risk of Profile Matching Attacks in Online Social Networks

Borga Haktan Bilen
haktan.bilen@ug.bilkent.edu.tr
Bilkent University
Ankara, TURKEY

Group 5

Zeynep Doğa Dellal
doga.dellal@ug.bilkent.edu.tr
Bilkent University
Ankara, TURKEY

Alper Bozkurt
alper.bozkurt@ug.bilkent.edu.tr
Bilkent University
Ankara, TURKEY

İzgi Nur Tamcı
nur.tamci@ug.bilkent.edu.tr
Bilkent University
Ankara, TURKEY

Gizem Gökçe Işık
gokce.isik@ug.bilkent.edu.tr
Bilkent University
Ankara, TURKEY

KEYWORDS

Privacy, Online Social Networks (OSN), Differential Privacy (DP), Graph, Attribute, Profile, Matching, Distortion, Attack, Mitigation, Performance, Comparison

1 INTRODUCTION

In the evolving domain of online interactions, social networks play a crucial role in the propagation and spreading of information and the maintenance of social connections. However, the vast amounts of data generated and stored within these platforms raise significant privacy concerns, mainly through the use of profile-matching attacks. These attacks aim to link discrete data pieces across multiple platforms to uncover or predict user identities, which could lead to privacy violations or malicious exploitation of private data. Online social networks (OSNs) can be categorized into two broad topics: structured and unstructured networks[1]. Structured networks are those where user relationships are explicitly defined and can be visually and functionally represented as graphs. These connections provide a framework that can be exploited through graph-based attacks to deduce user identities or attributes by analyzing how users are interconnected. To give an example of a structured online social network, we can mention the popular platform Instagram, which has a follower/following relation between users. Conversely, unstructured social networks do not rely on explicit relational data and typically operate through the aggregation of user attributes and behaviors, making them susceptible to attribute-matching attacks [1]. Here, attackers use distinct pieces of information across platforms, such as similar usernames, links to other accounts, or shared content, to link profiles and identify individuals [1]. The threat of profile matching has led to the development and application of various distortion techniques to protect sensitive user data. These techniques involve altering data in ways that make it difficult for attackers to accurately link profiles across different platforms without substantially impacting the utility of the data for legitimate or truthful purposes. Such techniques include data perturbation, noise addition, anonymization, and more sophisticated methods like tokenization and hashing, each serving to hide or alter data points to protect user privacy [2]. The importance of these distortion techniques cannot be understated, as they serve as the primary defense against the ever-increasing sophistication of profile-matching attacks. Their implementation helps maintain the integrity and confidentiality of user data, balancing the need for privacy with the functionality and utility of social networks. However, the effectiveness of these techniques varies, and their deployment must be carefully managed to maintain the user experience and the operational purposes of the data they protect [2]. To summarize, our topic is "Comparison and Analysis of Distortion Techniques in Terms of Mitigating the Risk of Profile Matching Attacks in Online Social Networks." Thus, this report investigates the nuances of these challenges, exploring the effectiveness of different distortion techniques comparably in mitigating the risks associated with profile-matching attacks in online social networks. By understanding the strengths and limitations of each technique, we can better protect user privacy while maintaining the robust functionality of these digital platforms.

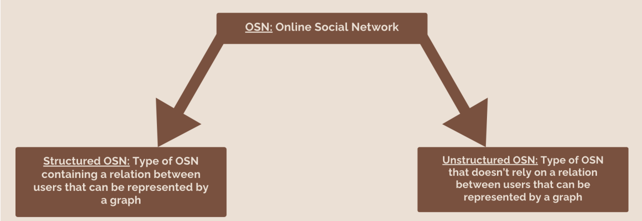

OSN: Online Social Network
Structured OSN: Type of OSN containing a relation between users that can be represented by a graph
Unstructured OSN: Type of OSN that doesn't rely on a relation between users that can be represented by a graph

Figure 1: OSN Categorization [2][1]

Finally, in the current literature, there doesn't exist an extensive survey regarding the performance comparison of the distortion techniques we tackle. Thus, we propose a **novelty** by creating a literature survey and empirical comparison regarding the performances of distortion techniques for profile-matching attacks on

, ,

OSNs (in essence, we tried to create a baseline comparison environment for every distortion technique, which is a lack in current works of literature). We also propose two system recommendations (one for each OSN type) to improve data privacy in online environments, per our novelty.

2 DETAILS OF SELECTED PROFILE MATCHING TECHNIQUES

2.1 Attribute Matching

Attribute matching, also known as record linkage, is a technique used in online social networks (OSNs) and various databases to identify records that refer to the same entity across different platforms. This process is critical for applications that aim to construct a detailed view of an individual's preferences and behaviors by aggregating data from multiple sources. The matching function exploits explicit user-provided information, such as usernames and biographical details, and inferred data from user activities, such as location stamps and timing of posts. In some cases, these mentioned attributes are named "Quasi-Idenfitiers." The technique depends on the identification of unique or highly indicative, or even unique attributes that remain consistent across platforms despite potential efforts by users to vary and perturb their information to maintain privacy or different online personas. The effectiveness of attribute matching is significantly enhanced by advanced algorithms that analyze these attributes for patterns of similarity, often employing methods such as machine learning to improve accuracy and reduce false matches. By integrating these diverse data points, attribute matching not only enables richer user profiles but also raises complex privacy and security implications, as it can unintentionally expose sensitive information or contribute to identity theft and other malicious activities. Throughout this essay, we will focus on security and data privacy implications caused by various attribute matching techniques. Furthermore, we will not focus on such attacking techniques because of the lack of sufficient empirical data on the performance of distortion techniques against attribute-matching attacks reinforced with machine learning algorithms. Additionally, we will assume the setting as an unstructured online social network throughout the discussions for attribute matching attacks. For the distortion techniques against attribute-matching attacks, we chose to focus on the following:

- Data Perturbation
- Noise Addition
- Anonymization
- Tokenization
- Hashing
- Suppression
- Generalization

Details regarding these topics will be discussed in further subsections. The main criterion for selecting these techniques is the wide availability of resources regarding the technical details and performances of these techniques. As an additional discussion, we diminished the initial large pool of distortion techniques against attribute-matching attacks because of either availability problems regarding the resources or the lack of uniform testing environments, in some researches performance of techniques is evaluated over a small set of users while in the others different distortion technique measured over a whole online social network (such as Instagram). This caused a varying environment for comparison, which would result in a biased outcome. [3] [4]

2.1.1 *Data Perturbation.*

Data perturbation is a fundamental approach within areas including privacy-preserving data mining, data analysis and related fields. It utilizes methodologies which are from disciplines such as statistical disclosure control and statistical databases. It involves modifying the original dataset to release a perturbed version for analysis while preserving privacy. A key point in data perturbation is the balance between privacy and data utility. On the other hand, perturbation techniques must ensure that the original data cannot be sufficiently reconstructed to reveal sensitive information.[5] In addition to this, they must allow meaningful patterns and insights to be extracted from the perturbed data.

Various techniques used in the scope of data perturbation, including additive, multiplicative, matrix multiplicative, k-anonymization, micro-aggregation, categorical data perturbation, data swapping and resampling. [5] Despite their differences, these techniques share the common goal of preserving sensitive information in datasets, making them essential tools in modern data privacy efforts.

2.1.2 *Noise Addition.*

Noise addition serves as a crucial strategy in achieving the balance between privacy and utility, masking numerical attributes to prevent inference attacks and reconstruction of sensitive data. By adding noise addition as a perturbation methodology, datasets can maintain confidentiality while maintaining their utility for consumers. [6]

Noise addition techniques, comprising additive and multiplicative noise, are fundamental in protecting personal data. Additive noise involves introducing random errors into the original data. Four primary procedures have been established: uncorrelated noise addition, correlated noise addition, noise addition with linear transformation, and noise addition with non-linear transformation. [7] Uncorrelated noise addition involves independently adding noise to each attribute, with the magnitude of noise determined by the standard deviation. On the other hand, correlated noise addition generates noise based on an error matrix, resulting in higher analytical predictability. Noise addition with linear transformation aims to maintain the sample covariance matrix of transformed attributes while non-linear transformation techniques address the limitations of the former, although requiring significant expertise and time. Multiplicative noise emerges as a solution to the challenge of constant variance in additive noise.

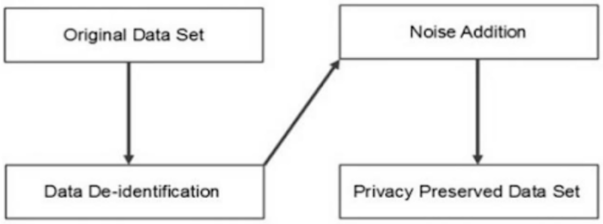


Figure 2: Generalized Data Privacy with Noise Addition [8]

2.1.3 *Anonymization.*

Anonymization is a crucial process employed when releasing data to prevent the disclosure of sensitive information about individuals. There are three main types of information disclosure that anonymization seeks to reduce: identity, attribute, and inference.[9] Identity disclosure occurs when anonymization techniques are insufficient, allowing for re-identification of specific records within the anonymized dataset. [10] For example, a simple anonymization process that replaces identifiable information with sequential numbers could lead to identity disclosure if the patterns are easily predictable. Attribute disclosure occurs when new information about individuals is unintentionally revealed. For instance, anonymized employee records may unintentionally disclose sensitive information, such as age-related additional incentives, which can be inferred based on other known attributes. Inference disclosure occurs when adversaries can derive confidential information by correlating anonymized data with other datasets without directly revealing sensitive information. As anonymization efforts increase, the practical value of the data decreases. [11]

2.1.4 *Tokenization.*

Tokenization is crucial in the information retrieval process; it plays an important role in segmenting text into distinct units, which are known as tokens. These tokens contain words, numbers, and characters, and their extraction is important for subsequent analysis. This process involves assessing the frequency of each token within the input documents, promoting a deeper understanding of the text. [12]

The tokenization process is described in several phases where documents are collected for word extraction. Later, infrequent words are eliminated, enhancing the efficiency of the tokenization process. Stop word removal follows, targeting redundant English words that offer minimal value in information retrieval. This phase not only reduces the size of indexing files but also enhances overall efficiency, contributing to the effectiveness of the tokenization process. [13]

2.1.5 *Hashing.*

Hashing plays a crucial role in ensuring data privacy and security. Hashing involves the transformation of input data of arbitrary size into a fixed-size string of characters, known as a hash value or hash code. One of the key characteristics of hashing is its deterministic nature, meaning that the same input will always produce the same hash value. This property enables efficient storage and retrieval of data, as it allows for rapid comparisons between hash values, not the original data itself. Commonly used hashing algorithms are MD5 (Message Digest Algorithm 5) and SHA-1 (Secure Hash Algorithm 1). [14]

Hashing methods are widely utilized in various applications, ranging from data verification and authentication to password storage. In password storage, for instance, hashing ensures that passwords are securely stored without being directly accessible in their plain-text form. When a user inputs their password, the system hashes the input and compares it with the stored hash value. If the two hashes match, authentication is successful without the need to store the actual password. In addition to password protection, hashing can also be used to protect other sensitive information, such as address information.

2.1.6 *Suppression.*

Suppression techniques are essential in privacy-preserving data mining, particularly in situations where extracting valuable understanding is necessary while maintaining the confidentiality of individuals' information. With the increasing availability of personal data and the advancement of data mining algorithms, concerns regarding privacy violations have grown significantly. Techniques such as classification, k-anonymity, association rule mining, and clustering have emerged as potential solutions to address these concerns. [15]

An illustrative instance of suppression could be the airport security of the scenario where passenger information records contain sensitive personal data such as names, passport numbers, demographic details, and flight information. To protect individual privacy, suppression techniques can be applied to de-identify the dataset by removing unique identity fields. However, even after such measures are taken, there remains the risk of identifying individuals through other available data attributes. Thus, the development of effective data mining algorithms for privacy preservation becomes crucial. This emphasizes the importance of ongoing research in this field, aiming to compare and contrast different approaches and develop frameworks for preserving privacy while extracting valuable understandings from data. Moreover, intentional distortion of information through suppression can lead to artificial inferences that are inaccurate and serve specific purposes with the reported values. Conversely, suppression may not be suitable when data mining necessitates full access to sensitive values. In such cases, limiting the identity link of a record may be a preferred method for preserving privacy. [16]

2.1.7 *Generalization.*

Generalization techniques are important in various machine learning tasks such as classification, regression, and clustering. This concept is aimed at ensuring that the model can capture fundamental patterns and make reliable predictions on new instances. In classification tasks, a well-generalized model can accurately classify new instances into defined categories based on the patterns learned from the training data. In regression tasks, a generalized model can effectively predict continuous outcomes for hidden data points. [17] Generalization involves not only selecting appropriate algorithms but also processing the data and evaluating the model's performance to ensure that it can generalize well beyond the training data.[18]

2.2 Graph-Based Attacks

Graph-based attacks in online social networks focus on exploiting the social graph—representing users and their connections—to link sanitized (anonymized) and desanitized (original) data graphs. These attacks attempt to re-identify anonymized nodes by exploiting graph structures and user attributes. By analyzing the connections and similarities between nodes in the sanitized graph and comparing them with publicly available or desanitized graphs, attackers can infer identities or sensitive attributes that were meant to be protected. This technique leans on the inherent relational data within social networks, where even indirect connections, such as friends of friends, can provide enough information to breach privacy. The effectiveness of these attacks increases with the availability of auxiliary data and the improvements of algorithms used to analyze graph topologies and attribute correlations, potentially causing significant risks to user anonymity and data confidentiality in online social networks [19]. We could mention their primary use case settings for comparing graph-based attacks to attribute-matching attacks. The graph-based attack techniques are evaluated under structured online social networks, while attribute matching techniques are looked at under unstructured online social networks. The leading cause of this difference originates from the fact that graph-based techniques utilize connections between users rather than solely focusing on the attributes of individual records. Thus, even though attribute matching and graph-based techniques use comparable and similar algorithms underneath, the settings they operate in are different. Throughout this essay, we will focus on the implications of security and data privacy caused by various graph-based attacking techniques. For the distortion techniques mitigating the graph-based attacks, we chose to focus on are as follows:

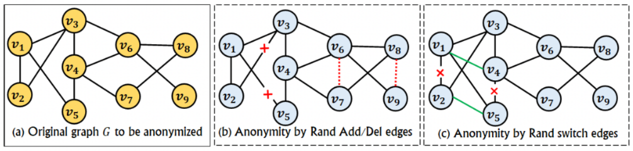
- Modification Method
- Clustering Method
- Privacy-Aware Graph Computing-Based(PAGC) Method
- Differential Privacy-Based
- Hybrid

Details regarding these topics will be discussed in further subsections. The main criterion for selecting these techniques is the wide availability of resources regarding the technical details and performances of these techniques. A similar comment can be made, as attribute matching attacks, for our selection process, in which we reduced an initial large pool of possible distortion techniques for graph-based attacks. We also decided not to include artificial intelligence-based graph-altering methods because we already investigate computer vision-based attacks and distortion techniques specifically. Thus, this discussion would merely act as a duplication of the same techniques applied to different scenarios, which is out of the scope of this essay. Finally, the hybrid method is added merely to observe whether the heuristic of combining multiple methods in one attack is actually effective. [19]

2.2.1 ***Modification Method.*** Graph modification methods play a critical role in preserving privacy in structured online social networks. These methods involve strategic alterations and modifications to the graph's structure, which represents the network of user connections, to prevent malicious entities, namely attackers, from accurately deducing personal or sensitive information about the network's users. The primary aim of graph modification is to conceal the original, often sensitive, structure of the social network. This is achieved by adding or deleting nodes and edges or by altering their attributes in a way that the modified graph remains practically useful while significantly reducing the risk of privacy breaches. In essence, the main goal of this distortion technique is to modify the topological structure of the graph while preventing loss of utility. The two main categories that are possible to mention under graph modification methods are as follows:

Adding and Deleting Edges: By introducing new edges between nodes or removing existing ones, the apparent relationships and interaction patterns among users are changed. This can prevent attackers from confirming hypotheses about social connections or inferring new connections based on known patterns. Another type of possible edge modification is rewiring already existing edges. However, it should be considered that these types of modifications can heavily affect the utility of graphs representing the social connections within the platform. [19]

Node Modification: Modifying nodes involves changing the node's attributes or its entire identity within the graph. This might include changing demographic details, user behaviors, or any other quasi-identifiers that could be used for linking profiles across platforms or within the same platform. This type of modification also prevents users' data from attribute matching attacks, which focuses on identifying data other than the connections between users. [19]



(a) Original graph G to be anonymized (b) Anonymity by Rand Add/Del edges (c) Anonymity by Rand switch edges

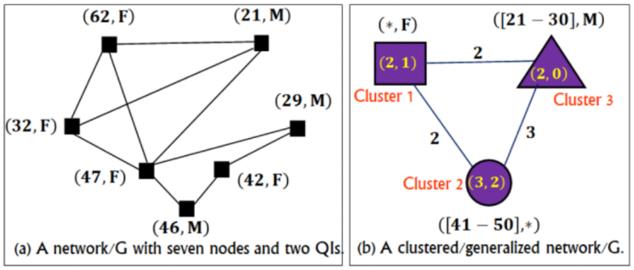
Figure 3: Graph Modification by Adding and Deleting Edges [19]

These methods are vital for maintaining the utility of the OSN while protecting user privacy. They balance between altering the graph enough to prevent accurate profile matching and maintaining enough structure to keep the network functional and informative. Effective implementation of graph modification methods can significantly reduce the risk of privacy attacks without degrading the user experience or the analytical value of the data for legitimate purposes. Additionally, all the mentioned graph modification methods can also be utilized by adding dummy nodes (vertices) to the graph, which is a better approach for preventing the loss of utility compared to altering original connections. [19]

2.2.2 ***Clustering Method.*** Graph clustering, also known as graph generalization, clusters nodes, and edges into super nodes and super edges, respectively, to protect sensitive user information from being exposed through graph-based attacks. The essence of graph clustering is to create a coarser version of the original graph where multiple nodes are grouped into a single super node. This approach effectively reduces the granularity of the graph, making it difficult for attackers to identify individual users or distinguish (de-anonymize) specific user attributes [19]. By clustering similar nodes together, the method masks the precise details of individual connections, thereby preserving the privacy of the users' identities and their interactions within the network.

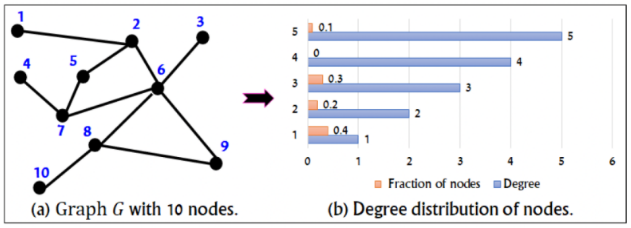
Regarding how clustering algorithms work, they rely on grouping nodes that share similar attributes or connection patterns. Thus, not only the characteristics of individual nodes are essential but the connections and degrees of nodes are also crucial. This not only confuses the trail for potential attackers but also maintains the utility of the graph for legitimate analysis for scientists, like community detection or network structure analysis [19]. Additionally, the super edges, connections between super nodes, reflect the overall connectivity pattern among clusters, not individual relationships.

In terms of graph publishing, clustering methods can be tailored to the needs of the researchers. For example, in a social network graph, users within the same geographical area or with similar interests might be clustered together to form a super node, hiding



(a) A network/G with seven nodes and two QIs. (b) A clustered/generalized network/G.

Figure 4: Graph Clustering Example [19]



(a) Graph G with 10 nodes. (b) Degree distribution of nodes.

Figure 5: Overview of a Degree Computation from a Graph [19]

individual user details but allowing analysis of broader community trends. So, researchers whose concern is the analysis of some specific interests can use these anonymized graphs without losing none-to-zero utility. This is one of the key differences between the clustering method and the graph modification method. For limitations, while clustering enhances privacy, it must be carefully designed to avoid under or over-clustering. Over-clustering may lead to a significant loss of useful information (utility), potentially rendering the graph useless for certain types of analyses. Under-clustering, on the other hand, might not provide enough privacy protection, leaving the data vulnerable to inference attacks.

Graph clustering methods are typically evaluated based on their ability to balance the trade-off between data utility and privacy. Effective clustering algorithms are those that maximize the entropy (a measure of information rate over a specific channel in which certain pairs can be confounded) [20] in each super node while maintaining enough structural information in the graph for analysis purposes.

2.2.3 ***Privacy-Aware Graph Computing (PAGC) Method***. Privacy-aware graph Computing (PAGC) methods are advanced privacy-preserving techniques that utilize computational algorithms to process and analyze graph data while maintaining the privacy of the information embedded in the graph [19]. These methods are crucial in structured OSNs where the explicit relationships between users can be exploited to infer personal information. PAGC involves the application of computational techniques that are designed to function under privacy constraints. This may include modifying the graph's topology or employing algorithms that ensure privacy through differential privacy, secure multi-party computation, or homomorphic encryption. The goal is to perform necessary computations like community detection, influence measurement, or path analysis without compromising the privacy of the individual nodes (users). In other words PAGC methods aim to reveal interesting characteristics of graphs rather than perturbing the graph itself. This can be thought of as extracting various statistics (for instance, degree computation) from the graph for publishing or preserving the data. However, it is not a common approach to use PAGC methods to store social network data in databases or data warehouses because it completely negates the need for a structural scheme for online social networks to hold the connection data between users [19].

As far as PAGC algorithms, we can leverage differential privacy techniques by adding noise to the outputs of queries performed on the graph data, such that the presence or absence of a single node (which is the user) does not significantly alter the output, thereby preserving the individual's privacy. Thus, also prevents the OSN from using inference attacks. For statistics extraction, it is also possible to break the graph into smaller, less informative subgraphs that can be processed separately. This limits the amount of information an attacker can infer from any single computation [19].

Implementing PAGC methods involves a careful balance between privacy and utility, similar to other distortion techniques. Too much noise or overly aggressive decomposition can easily render the data useless, while too little may not sufficiently protect privacy. Furthermore, these methods often require more complex computations, which can be resource-intensive.

2.2.4 ***Differential Privacy-Based (DP) Method***. Differential Privacy-Based (DP) methods provide a robust framework for preserving privacy in structured OSNs by ensuring that the removal or addition of a single individual's data does not significantly affect the outcome of queries made against the database [19]. This is achieved by injecting a carefully calibrated amount of random noise into the data or query results, thus guaranteeing privacy while still allowing for the utility of the data. In a sense, we are making small alterations to the overall graph while preserving key statistics regarding the graph structure and the node attributes. Differential privacy introduces the concept of the Privacy Budget, which quantifies how much information an individual query reveals. Each query is deducted from the previously decided budget, with more costly queries consuming more. Conversely, we can also introduce a Privacy Loss Budget, which signifies the maximum amount of information leakage (in essence, exposure of sensitive data) that can be introduced to the output dataset (refer to Fig. 6, the epsilon represents the privacy loss budget and G1 and G2 represents two graphs that differ by just one node). The method ensures that the statistical noise added is proportional to the sensitivity of the queries, thereby masking the presence or absence of any single user's data.

In structured OSNs, where data is represented as graphs, DP can be applied in various ways, both in centralized or decentralized settings. One common approach is by adding noise to the properties of the graph such as node degrees or the existence of specific edges. This helps prevent attackers from accurately inferring relationships and characteristics of the individuals within the network. This type

$$\frac{Pr(\Im(G_1) \in S)}{Pr(\Im(G_2) \in S)} \leq exp(\epsilon)$$

Figure 6: General DP Model for Any Anonymization Algorithm Using Privacy Loss Budget [19][21]

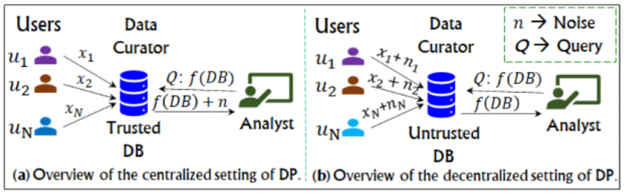


Figure 7: Overview of Two Most Common DP Settings [19]

of distortion can also prevent data on OSN against attribute matching attacks. DP techniques can also include randomized response strategies in the querying process and the Laplace or exponential mechanism to add noise to the data outputs. For instance, when calculating the degree of connectivity of a node, noise is added to the degree count before it is published or used in further analysis. Regarding these, the primary benefit of DP is its strong theoretical guarantees of privacy, which hold under a broad set of conditions and adversarial models [19]. However, a significant challenge is balancing privacy with data utility, as excessive noise can diminish the usefulness of the data. Additionally, implementing DP in a graph model requires careful consideration of graph-specific characteristics, such as clustering coefficients and community structures, which might be distorted by the added noise. However, with the introduction of a budgeting mechanism, it is possible to optimize the utility loss while mitigating the risk of profile-matching attacks.

2.2.5 ***Hybrid Method***. Hybrid methods in privacy-preserving graph anonymization apply a combination of techniques to enhance the effectiveness of privacy protection in structured online social networks. These methods leverage the strengths of multiple individual techniques, compensating for the limitations of each through the integration of another distortion technique, providing a more robust defense against privacy attacks. In other words, these methods try to leverage the synergy between different distortion techniques to increase data privacy with minimal utility loss.

As mentioned, hybrid methods typically integrate two or more privacy-preserving strategies, such as graph modification, differential privacy, and clustering. For instance, a common hybrid approach might involve the application of differential privacy to the results of a graph clustering method. This combination ensures that the anonymized graph not only masks the identities through clustering but also protects against inferential attacks by adding noise to the aggregated data. The main advantage of using hybrid methods is their ability to provide higher levels of privacy without significantly compromising the utility of the data. Combining methods makes it more challenging for adversaries to exploit the weaknesses inherent in a single technique. For example, while differential privacy adds noise that can reduce data utility, clustering can preserve important structural properties, allowing for meaningful analysis without revealing sensitive information. This property of hybrid methods also makes it possible for a distortion scheme to be tailored to the needs of a researcher (for publishing reasons) or an environment to protect sensitive data from a very specific attack.

On the other hand, implementing hybrid methods comes with its set of challenges, particularly in balancing the trade-offs between privacy and utility. It requires fine-tuning of parameters to ensure that the added noise does not overwhelm the valuable information or characteristics in the data [19]. Moreover, the complexity of hybrid methods can increase the computational overhead and require more sophisticated algorithms to effectively manage the interactions between different techniques. Thus, it is accurate to say that applying multiple methods in combination requires more know-how regarding the synergy of each distortion technique against each other.

2.3 Face Recognition/Computer Vision-Based Attacks

Face recognition or computer vision-based attacks in online social networks exploit the vulnerabilities inherent in facial recognition systems or leverage the affluence of image data to deanonymize users. These attacks utilize sophisticated techniques to match facial data across various platforms or datasets, aiming to uncover the identities of anonymized users or to create fraudulent or basically wrong and harmful identities. Such attacks involve collecting face images from different social platforms where users might not even be aware of being targeted. The attackers then use advanced computer vision techniques and algorithms to enhance, analyze, and compare these images against other datasets. This can include using machine learning models to improve the accuracy of matching faces under varying conditions, such as different lighting, angles, or facial expressions. The ultimate goal of these attacks is to breach data privacy by linking a person's separate online personas or social network accounts, revealing their activities across different platforms, or even impersonating them in criminal activities. As these techniques continue to evolve, they propose significant challenges to maintaining user privacy and security on digital platforms. Additionally, because the domain of machine learning and computer vision is a newly emerging and fast-developing area, further empirical performance analysis regarding these attacks and the distortion techniques against these attacks are still a requisite in the data privacy domain. We will strictly focus on face recognition or computer vision-based profile linkage attacks in the setting of unstructured online social networks. This essay will focus on the security and data privacy implications of various computer-vision/face recognition-based attacking techniques. For the distortion techniques against graph-based attacks, we chose to focus on the following: [22]

- K-Anonymity
- Image Perturbation

Details regarding these topics will be discussed in further subsections. The main criterion for selecting these techniques is the wide availability of resources regarding the technical details and performances of these techniques. A similar comment can be made, as attribute matching attacks and graph-based attacks, for our selection process, in which we reduced an initial large pool of possible distortion techniques for computer vision algorithms and attacks based on them.

2.3.1 ***K-Anonymity.*** The concept of anonymity was first introduced by Sweeney (2002) to address the issue of protecting individuals' identities in a dataset. The principle behind K-Anonymity is to ensure that each data record is indistinguishable from at least k-1 others with respect to certain "quasi-identifier" attributes. This technique is important in scenarios where data must be shared or published without revealing sensitive individual information [23]. For example, in a medical dataset, attributes such as age, ZIP code, and gender could be considered quasi-identifiers. By ensuring that each combination of these attributes appears at least k times in the dataset, anonymity can effectively prevent the re-identification of individuals. We can also use k-anonymity for these examples, but what we are really examining is k-anonymity for images. We do this using a k-1 image decoy.

A study by Aggarwal and Yu (2008) has various models that improve the k-anonymity approach, making it applicable to a wider range of data types and more resilient against various types of data mining attacks. The model's effectiveness is dependent on the selection of k and the specific attributes considered as identifiers [24]. However, a notable limitation of k-anonymity is its vulnerability to homogeneity, and background knowledge attacks, where an attacker might have sensitive information from the k-anonymized data if subsets within the data are not diverse enough or external information is available.

For comparing performances in the following sections of this research we selected to focus on the following specific K-Anonymity technique:

- K-Anonymity with image decoy

2.3.2 ***Image Perturbation.*** Image Perturbation techniques involve the modification of image data to prevent unauthorized access to or recognition of the information. It refers to the modification of image data in a way that maintains the utility of the data for legitimate purposes (like facial recognition for authentication) while preventing misuse in identity theft or unauthorized tracking. Techniques such as adding noise, blurring, and other forms of digital alteration fall under this category.

A study by Newton et al. (2005) involves the use of k-same face de-identification, which perturbs facial images so that they are less likely to be original individuals but remain useful for demographic analysis. This method provides a balance between privacy and utility by ensuring that face images in a dataset cannot be easily linked back to individuals.[25]

A study by Gross et al. (2006) presents a system for automatic face anonymization within videos. Their system applies real-time perturbation techniques to face images, significantly reducing the risk of personal identification while maintaining enough detail for audience analysis. [26]

A study by Moosaavi et al. (2017) introduced the notion of universal adversarial perturbations, which are crafted perturbations that can mislead a wide range of computer vision models. This discovery not only emphasizes the vulnerabilities of deep learning models in image recognition tasks but also highlights a method for protecting image data from vulnerabilities. [27]

For comparing performances in the following sections of this research we selected to focus on the following specific image perturbation technique:

- **Gaussian blurring [28]**
  Explanation: Gaussian blurring is an image processing technique that uses a Gaussian function to convolute an image. The Gaussian function is applied to calculate the transformation magnitude applied to each pixel in the image. The function takes into account the horizontal (x) and vertical (y) distances from the center (0,0), within a circular recognition domain. The values coming from these coordinates are then used to construct the convolution matrix that is applied to the original facial image. After the convolution process, each pixel is assigned a new value, which is the weighted average of its neighboring pixels. The original pixel, having the closest coordinates to the center and thus the highest Gaussian value, is assigned the highest weight. The surrounding pixels contribute to the averaged pixel value based on their distance from the center. This process results in a low-pass filter output, producing a blurred image with minimized clear edges of the facial features. The blurred edges play a pivotal role in distorting facial recognition algorithms. In most Online Social Networks (OSNs) such as Instagram, face detection models like the Haar Cascade are employed. This algorithm classifies face-like objects based on the intensity and edges of the pixels. Gaussian blurring serves as an effective distortion technique in this initial step, as no apparent edges are detectable to a model that is not specifically trained with blurred images. This aids in protecting the user from being identified. Following detection, facial analysis is conducted where the facial landmarks are extracted and analyzed using methods like the Constrained Local Model (CLM). This method captures the variation in pixels within patches of the picture and associates them with aligned facial features. These methods yield a unique and distinct faceprint of the user, which is then compared against a database of known faces. The distorted faceprint significantly reduces the chances of matching to these known faceprints, thereby serving to de-identify users' faces. The efficacy and limitations of this technique will be discussed in further sections.
- **Differentially Private Face Pixelation [29]**
  Explanation: The proposed algorithm commences with a process known as pixelization applied to the input image. This technique involves the substitution of each individual pixel in an image with a larger block of pixels, all of which share the same value. Typically, this value corresponds to the average color of the original pixels. The consequence of this process is a reduction

in the image's resolution, thereby making it more challenging to discern fine details for face detection and facial analysis models explained above. However, pixelization is a deterministic process, always producing identical results given the same input. This predictability, however, can potentially be exploited by neural networks to re-identify faces in pixelized images, overlooking k-same anonymity principle in matching attacks. Therefore, to introduce an element of randomness and enhance protection against re-identification, the algorithm employs Laplace perturbation on the pixelized image. The Laplace distribution, a probability distribution commonly used in differential privacy, is used to introduce noise into the data. This ensures that the output, in this case, the pixelized image, does not give away excessive information about any single user. The noise is tried to be included in a way that balances between privacy and utility which will be evaluated in further sections. The combination of pixelization and Laplace perturbation introduces a degree of randomness that serves to protect against re-identification.

- **Noise Addition [30]**
  Explanation: Noise addition is a widely used technique of processing images that obscure the face and makes it harder for FR models to extract features from facial analysis. However, noise addition can come with utility costs for human perception. Therefore, a newly proposed method, known as Differential Privacy of Landmark Positioning (DPLP), adds noise to specific, sensitive areas of face images to protect privacy. The noise is added in such a way that it is difficult for face recognition algorithms to correctly identify the face, but the changes are subtle enough to be almost imperceptible to humans. The DPLP method works by first using the Active Shape Model (ASM) algorithm to position the area of each face landmark. If a landmark overlaps a subgraph of the original image, then the subgraph is considered a sensitive area. This sensitive area is then treated as the seed for regional growth, following the Fusion Similarity Measurement Mechanism (FSMM). The privacy budget, which is a measure of the amount of privacy protection applied, is only allocated to the seed. The privacy budget is a measure of the amount of privacy protection that is applied to the data. The allocation of the privacy budget to the seed is where the noise addition comes into play. The noise is added to the seed area in a manner that provides privacy protection while preserving the utility of the data.The noise added to the image is designed to exploit the weaknesses of face recognition algorithms, and when it comes to black box attacks, where the model is not well known, implementing this method becomes more challenging which will be discussed in further sections.
- **Face Feature Space Perturbation [31]**
  Explanation: The Feature Space Adversarial Perturbation (FSAP) framework is a novel approach to face image de-identification. This method is conducted by introducing adversarial perturbations in the feature space of the image, which are designed to mislead deep neural networks used for automated face recognition. This is a significant departure from traditional methods of image de-identification, which typically involve pixel-level transformations such as blurring or masking. The FSAP framework includes a specially designed algorithm for generating these adversarial perturbations. The algorithm works by alternating the perturbation based on two loss functions: the ID loss and the attribute loss. The ID loss is related to the identity information of the individual in the image. The algorithm directs the adversarial noise towards the identity-related features of the image, effectively hiding this information from DNNs. The attribute loss, on the other hand, is related to other attributes of the individual that are not directly linked to their identity, such as their age or emotion. By minimizing this loss, the algorithm ensures that these attributes remain similar even after the introduction of the adversarial noise.
- **Privacy-Protective-GAN for Face De-Identification [32]**
  Explanation: Privacy-Protective GANs (PP-GANs) represent an innovative approach to face de-identification, addressing privacy concerns in the context of face recognition technologies. The central challenge lies in striking a balance between safeguarding privacy and maintaining visual quality. PP-GANs build upon the well-known Generative Adversarial Network (GAN) architecture. The generator creates de-identified face images by learning from random noise, while the discriminator distinguishes between real (original) and generated (de-identified) data. During training, the generator learns to create realistic face images by minimizing the difference between its generated samples and real face images. The verificator module ensures that the generated face still resembles the original person, incorporating domain knowledge (e.g., facial landmarks, identity-specific features). The regulator module allows users to customize the level of de-identification, balancing privacy preservation and visual realism. PP-GANs directly impact user privacy, achieving better privacy protection than traditional methods while preserving visual resemblance.

Table 1: Key Concerns of Each Distortion Technique for Computer Vision-Based Attacks

Topic	**Key Concern**
K-Anonymity	Privacy protection through indistinguishability
Image Perturbation	Enhancing image data security and privacy

3 PERFORMANCE COMPARISON OF DISTORTION TECHNIQUES

3.1 Comparison of Distortion Techniques for Attribute Matching Attacks

For Attribute Matching Attack Distortion Techniques, their effectiveness in preserving data utility while protecting privacy will be evaluated. Additionally, even though both recall and precision is going to be mentioned, for comparison reasons and system recommendation we will base the F1-Score. This is because of the fact that the experiments done using these distortion techniques are done on a very diverse set of online social networks. Thus, it wouldn't be an accurate choice to base our recommendations just by considering type-1 or type-2 error. We believe that harmonic recall and precision is more valuable. Here are how metrics will be used to assess for each distortion technique:

3.1.1 *Data Perturbation*.

F1-Score: Evaluate the harmonic mean of precision and recall after applying data perturbation. This metric considers both false positives and false negatives, making it suitable for assessing the overall performance of the perturbation technique. This metric is evaluated for two different data perturbation approaches named NOS2R and NOS2R over 10 different data sets. [33] Robustness: Measure the resilience of data perturbation against adversarial attacks aiming to reverse the perturbation and re-identify individuals. [34]

Sensitivity Analysis: This aspect evaluates how changes in the parameters of perturbation methods influence the performance of machine learning models. For instance, in the image domain, perturbation techniques like Noise, Brightness, and Contrast exhibit discernible patterns in performance variation as their parameters are adjusted. Typically, higher levels of perturbation result in decreased model performance across all perturbation types and tested models. This trend highlights the sensitivity of models to perturbation strength, indicating a diminished capability to accurately classify images as perturbation intensity increases. Similarly, in the audio domain, models show varying responses to changes in perturbation strength, with different perturbations such as White Noise, Compression, and Pitch demonstrating distinct effects on model performance.

Noise Tolerance: Noise tolerance refers to the ability of machine learning models to maintain performance in the presence of added noise or perturbations. In the audio domain, some perturbators, such as Clipping and Volume, showcase a linear or normal distribution in their impact on model performance. This suggests a certain level of tolerance to alterations in audio quality or volume. However, other perturbators, like Compression and White Noise, significantly impact model performance, especially at higher perturbation strengths. This indicates lower noise tolerance in these cases, as the models struggle to maintain accuracy when subjected to intense perturbations.

Overall Robustness: Robustness, in the context of data perturbation, encompasses the ability of machine learning models to sustain performance across diverse perturbation conditions. A robust model exhibits stable performance despite variations in perturbation strength and type. This is evaluated by considering both sensitivity analysis, which assesses the impact of perturbation parameters on model performance, and noise tolerance, which gauges the model's ability to maintain accuracy in the presence of added noise or perturbations. Ultimately, a robust model demonstrates consistent and reliable performance across a range of perturbation scenarios, making it suitable for real-world applications where data integrity may be compromised.

Robustness Categorization: Medium

Explanation: Data perturbation techniques, such as NOS2R and NOS2R, exhibit medium robustness. This conclusion is based on their ability to maintain a balance between preserving data utility and privacy (as indicated by the F1-score) and their resilience against attempts to reverse-engineer the perturbed data to re-identify individuals (as indicated by the robustness metric). While these techniques effectively perturb the data to protect privacy and maintain utility, they may still be vulnerable to sophisticated attacks targeting the perturbed data, hence the medium categorization.

3.1.2 *Noise Addition*.

F1-Score: Compute the F1-score to assess the balance between precision and recall after adding noise to the data. [35] The F1-score of noise addition represents the average performance across two distinct datasets.

Robustness: Evaluate the ability of noise addition to withstand attempts to reverse-engineer the original data from the noisy version. Noise addition robustness is evaluated in the context of training neural networks with standard Stochastic Gradient Descent (SGD) and Differentially Private SGD (DP-SGD). The robustness is measured against adversarial examples generated using techniques like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), as well as common corruptions like brightness, fog, and noise.

Sensitivity Analysis: FGSM Robustness Measurement: Sensitivity analysis is conducted by evaluating the model's accuracy as the strength of the adversarial perturbation (ϵ) increases. Both SGD and DP-SGD models show decreasing accuracy as the perturbation strength increases, indicating sensitivity to adversarial perturbations. However, the accuracy drop is more pronounced for models trained with DP-SGD, suggesting that the addition of noise during training may amplify sensitivity to adversarial examples.

PGD Robustness Measurement: Sensitivity analysis is also performed by comparing the accuracy of models under PGD attacks with varying strengths (ϵ) and number of steps. The results show that DP-SGD-trained models exhibit larger accuracy drops compared to SGD models for certain combinations of ϵ and steps, indicating increased sensitivity to adversarial attacks.

Noise Tolerance: Effect of Noise on Robustness: The study explores the effect of noise (σ) and clipping bounds (C) on the robustness of models against FGSM attacks. It is observed that increasing the noise level and clipping bounds decreases the robustness of DP-SGD-trained models, as indicated by larger accuracy drops under adversarial attacks. This suggests that while noise addition may enhance privacy, it can also compromise the model's ability to resist adversarial attacks.

Overall Robustness of Noise Addition: FGSM and PGD Results: The experiments demonstrate that models trained with DP-SGD exhibit decreased robustness compared to those trained with standard SGD, as evidenced by larger accuracy drops under both FGSM and PGD attacks. Additionally, DP-SGD trained models show decreased robustness to common corruptions compared to SGD models, indicating that the noise addition during training may not generalize well to handling various forms of data corruption. Overall, the noise addition robustness measurements suggest that while DP-SGD enhances privacy, it may come at the cost of decreased robustness to adversarial attacks and common corruptions. Sensitivity analysis reveals that the addition of noise during training may amplify the model's sensitivity to adversarial perturbations, highlighting the trade-off between privacy and robustness in differential privacy techniques.

Robustness Categorization: Low

Explanation: Noise addition techniques demonstrate low robustness due to their limited ability to withstand attempts to reverse-engineer the original data from the noisy version (as indicated by the robustness metric). Despite potentially preserving privacy to some extent, noise addition may not sufficiently obscure sensitive information from adversarial attacks or attempts to recover the original data. This vulnerability to attacks lowers the overall robustness of noise addition techniques.

3.1.3 *Anonymization.*

F1-Score: Calculate the F1-score to evaluate the overall effectiveness of anonymization in preserving privacy without sacrificing too much utility. The F1 score of the distortion technique anonymization was determined in this study by training Random Forest models on both the original Train dataset and the synthetic dataset generated by the SINTEZA engine. The original dataset consisted of clinical data from 5110 patients, with variables such as gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, smoking status, and stroke event. After randomly splitting the original dataset into Train and Test sets, the missing values in the BMI records were imputed using the KNNImputer class. Both datasets underwent oversampling of the minority class of the target variable (Stroke) using the SMOTE algorithm to improve model performance. Subsequently, Random Forest models were trained on both datasets without fine tuning or cross-validation. The F1 score, which considers both precision and recall, was then computed for each model using the Test dataset. The F1 score for the model trained with the original data was found to be 0.1889, while the F1 score for the model trained with the synthetic data was 0.1550. This comparison provides insight into the performance of the distortion technique of anonymization in preserving the underlying patterns and statistical information of the original dataset while ensuring patient privacy. [36]

Robustness: The focus of the paper on anonymization primarily centered on evaluating the robustness of the k-anonymity model, specifically through sensitivity analysis of the ARX k-anonymization algorithm. This approach enabled a comprehensive assessment of the model's effectiveness in reducing re-identification risk while maintaining data utility across various values of k.

Sensitivity Analysis: The importance of sensitivity analysis in data anonymization is highlighted in various studies, emphasizing the need for tailored approaches to balance privacy preservation with data utility. Different attributes exhibit varying sensitivity requirements, necessitating careful consideration during the anonymization process. For example, controlling the frequency of sensitive attributes ensures diversity within equivalence classes in anonymized datasets, contributing to enhanced privacy protection. The sensitivity of data, quantified by functions like S(f), reflects the inherent risk associated with data disclosure, with certain functions exhibiting higher sensitivity than others. Moreover, the contextual nature of data sensitivity suggests that personal information becomes sensitive based on its context, influencing the choice of data release options. The findings from the study demonstrate that sensitivity-based anonymization, combined with techniques like k-anonymization and l-diversity models, effectively mitigates re-identification risks while maintaining data quality. Comparisons of classifier performance further illustrate the trade-offs between anonymization and utility, with sensitivity-based approaches yield varying accuracies across different anonymization scenarios. [37] Ultimately, the analysis underscores the complexity of data anonymization in evolving information landscapes, highlighting the ongoing need for nuanced approaches to address privacy concerns without compromising data usability.

Noise Tolerance: The noise tolerance of anonymization techniques is crucial for preserving data privacy while maintaining data utility. Analyzing noise tolerance involves assessing how well an anonymization method can handle variations and perturbations in the data without sacrificing its effectiveness in protecting privacy. Anonymization techniques with high noise tolerance can effectively obscure sensitive information while preserving the overall structure and characteristics of the data. This means that even when subjected to various forms of noise, such as random perturbations or intentional attacks, the anonymized data remains sufficiently protected against re-identification while still being useful for analysis and other purposes. Assessing noise tolerance typically involves evaluating the impact of noise on the anonymized data's privacy protection and utility. Anonymization methods that can withstand moderate levels of noise without significantly compromising privacy or data quality are considered to have high noise tolerance. These methods often employ techniques such as generalization, suppression, and randomization to introduce controlled noise into the data while preserving its integrity. Research in anonymization often focuses on developing techniques that strike a balance between noise tolerance, privacy protection, and data utility. By enhancing noise tolerance, anonymization methods can better adapt to diverse datasets and real-world scenarios, ensuring effective data privacy in various applications, including network anonymization. [38]

Overall Robustness: The overall robustness of anonymization techniques seems to be medium to high. This conclusion comes from evaluations using metrics like the F1-score and sensitivity analysis. The k-anonymity model shows effectiveness in reducing re-identification risks while maintaining data utility, contributing to this robustness. Noise tolerance, another key aspect, ensures that anonymized data remains resilient to variations and perturbations without sacrificing privacy or utility. Anonymization methods with high noise tolerance can hide sensitive information while keeping the data's structure intact, enhancing overall robustness. However, achieving robust anonymization requires careful consideration of factors like sensitivity requirements and trade-offs between privacy and utility. Ongoing research is needed to further enhance robustness and address evolving privacy concerns.

Robustness Categorization: High

Explanation: Anonymization techniques, such as k-anonymity models, exhibit high robustness. This conclusion is drawn from their effectiveness in reducing re-identification risks while maintaining data utility (as indicated by the F1-score) and their ability to withstand attempts to reverse-engineer the anonymized data (as indicated by the robustness metric). These techniques effectively balance privacy preservation with data utility and demonstrate resilience against re-identification attacks, thus achieving high robustness.

3.1.4 *Tokenization*.

F1-Score: Compute the F1-score to assess the effectiveness of tokenization in preserving privacy and data utility. [39] The performance of five tokenization methods (W1, W1p, C3, C4, C5) was evaluated based on nine criteria within each method. To derive the overall effectiveness of these methods, the F1-score was calculated for each criterion, and the results were averaged across all criteria. The resulting ultimate average F1-score, obtained by summing and averaging the scores across all methods and criteria, provides a comprehensive assessment of the tokenization methods' performance.

Robustness: Evaluate the resilience of tokenized data against attacks aiming to reverse-engineer the original data from the tokens. [40]

Sensitivity Analysis: The scale-space tokenization method offers a significant advancement in sensitivity analysis by effectively balancing the robustness and utility of Vision Transformer (ViT) models. Through the incorporation of scale-space patch embedding and positional encoding, the method enhances the model's resistance to adversarial and out-of-distribution perturbations. Experimental validation on benchmark datasets such as CIFAR10/100 and ImageNet-1k reaffirms the method's efficacy in preserving model robustness while ensuring practical utility.

Noise Tolerance: This approach substantially enhances noise tolerance by leveraging shape bias and scale-space representation. Larger patch-sized ViT models with increased shape bias demonstrate superior resilience to various perturbations, including pixelation, fog, rain, and snow. By employing scale-space representation, the method effectively suppresses fine-scale details while retaining essential shape information, thereby bolstering noise tolerance in Tokenization models.

Overall Robustness: The introduced method significantly bolsters the overall robustness of ViT models against adversarial perturbations and common corruptions. Future research avenues may explore further enhancements in noise tolerance and sensitivity analysis within the Tokenization framework. Extending robustness improvements to encompass other perturbation types, such as natural adversarial examples and naturally occurring distributions, holds promise for enhancing noise tolerance in Tokenization models. The method's emphasis on shape bias and scale-space representation underscores its efficacy in bolstering noise tolerance and sensitivity analysis, thereby enhancing the overall robustness of Tokenization models.

Robustness Categorization: Medium

Explanation: Tokenization methods demonstrate medium robustness due to their effectiveness in preserving privacy and data utility (as indicated by the F1-score) and their ability to resist attempts to reverse-engineer the tokens back to the original data (as indicated by the robustness metric). While tokenization enhances privacy protection and maintains data utility, it may still be susceptible to certain attacks targeting the tokenized data, leading to a medium categorization of robustness.

3.1.5 *Hashing*.

F1-Score: Calculate the F1-score to evaluate the balance between precision and recall after applying hashing to the data. The F1 scores for the hashing algorithms A-hash, D-hash, W-hash, and P-hash were averaged to provide an overall assessment of their performance. This involved summing the F1 scores obtained for each algorithm across different distance thresholds and then averaging these values. The resulting average F1 scores were used to compare the overall effectiveness of the hashing methods. [14]

Robustness: Assess the cryptographic strength of hashing algorithms used and their resistance against hash reversal attacks.

Sensitivity Analysis: The introduced hashing schemes demonstrate significant improvements in sensitivity analysis, particularly in robustness and security features. Analytical expressions derived using differential entropy as a metric facilitate the assessment of security in hashing, ensuring a thorough examination of feature extraction stages. The schemes exhibit resilience to moderate filtering, compression operations, and common geometric manipulations, up to 10 degrees of rotation and 20 percent cropping, indicating a comprehensive sensitivity analysis approach.

Noise Tolerance: The hashing schemes showcase commendable noise tolerance, as evidenced by their ability to withstand moderate filtering and geometric operations without compromising robustness or security. The methods effectively identify and mitigate malicious manipulations, such as cut-and-paste editing, while preserving the integrity and discriminative capabilities of the hashed images. Differential entropy analysis confirms the schemes' noise tolerance, providing insights into their ability to maintain stability and security in the presence of varying levels of noise and perturbations.

Overall Robustness: The developed image hashing algorithms offer a balanced blend of robustness and security, catering to diverse applications such as authentication, watermarking, and image databases. Through comprehensive sensitivity analysis and differential entropy metrics, the schemes demonstrate advanced robustness against estimation and forgery attacks, ensuring a reliable representation of data for various practical scenarios. The incorporation of shape bias and scale-space representation enhances overall robustness, underscoring the schemes' effectiveness in preserving image integrity and security across different noise levels and manipulation types.

Robustness Categorization: Medium

Explanation: Hashing algorithms exhibit medium robustness, considering their ability to balance robustness and security features (as indicated by the robustness metric) and their effectiveness in preserving data integrity against attacks. While hashing provides a reliable representation of data and offers resistance against reverse-engineering attacks, it may still face challenges in maintaining security under certain adversarial scenarios, leading to a medium categorization of robustness.

3.1.6 *Suppression*.

F1-Score: Compute the F1-score to assess the overall effectiveness of attribute suppression in preserving privacy and data utility. The F1-score values for different methods were averaged to assess their performance in jamming suppression. Despite the focus on jamming recognition, an average F1-score specifically for suppression was derived by averaging the F1-scores of different methods across all instances of jamming. This involved summing up the F1-scores for suppression across all method-jamming combinations and then dividing by the total number of combinations. The resulting average F1-score provides an overall assessment of the effectiveness of the methods in suppressing jamming interference. [41]

Sensitivity Analysis: The technique effectively mitigates the risk of privacy breaches even when the suppressed data undergoes minor changes. By suppressing values that could potentially lead to the identification of individuals, the technique remains robust in sensitivity analysis. It can withstand variations in the suppressed data while still ensuring privacy, thereby reducing the likelihood of privacy risks.

Noise Tolerance: The suppression distortion technique enhances noise tolerance by effectively handling variations in the suppressed data caused by noise or random fluctuations. Despite inherent noise in the data, the technique maintains privacy protection intact. By suppressing sensitive information and distorting released data, it ensures that privacy is preserved even in the presence of random variations, thus enhancing its robustness in noise tolerance.

Overall Robustness: The technique demonstrates overall robustness by effectively thwarting privacy breaches across different types of queries and inference attacks. By systematically suppressing information that could reveal personally identifiable information (PII), it strengthens privacy protection in diverse scenarios. The technique's ability to withstand various attacks and scenarios underscores its effectiveness in ensuring privacy preservation in real-world applications.

Robustness Categorization: High

Explanation: Suppression techniques demonstrate high robustness due to their ability to effectively mitigate privacy breaches (as indicated by the F1-score) and maintain privacy protection in the presence of noise or variations in the suppressed data (as indicated by the noise tolerance metric). These techniques systematically suppress sensitive information and ensure privacy preservation across different scenarios, thus achieving high robustness.

3.1.7 *Generalization*.

F1-Score: Calculate the F1-score to evaluate the balance between precision and recall after applying generalization techniques.

The average F1 score was calculated based on the F1 scores obtained from testing various methods on the model generalization ability testing dataset. These methods include word2vec-BiLSTM-CRF, word2vec-BiGRU-CRF, word2vec-CRF, FastText-BiLSTM-CRF, FastText-BiGRU-CRF, FastText-CRF, BERT-BiLSTM-CRF, BERT-BiGRU-CRF, and BERT-CRF. The average F1 score provides an overall assessment of the performance of these methods in terms of their model generalization ability. [42]

Sensitivity Analysis: The generalization distortion technique exhibits robustness in sensitivity analysis by effectively handling variations in the data caused by the distortion process. Despite the transformation applied to the data for generalization purposes, the technique ensures that the model's performance remains stable across different scenarios. It can withstand changes in the data distribution induced by the generalization process, thereby maintaining consistent performance levels. [43]

Noise Tolerance: The technique enhances noise tolerance by mitigating the impact of noise introduced during the generalization process. Despite potential noise in the data resulting from the distortion technique, the technique maintains robustness by ensuring that the model's performance is not significantly affected. It can effectively filter out irrelevant information and focus on capturing essential patterns in the generalized data, thereby enhancing noise tolerance and preserving model robustness.[43]

Overall Robustness: The generalization distortion technique demonstrates overall robustness by effectively balancing the trade-off between model performance and generalization capability. It ensures that the model remains robust even after the data is generalized, thereby improving its ability to generalize to unseen data while maintaining stable performance in real-world scenarios. The technique's ability to preserve model robustness in the face of data generalization underscores its effectiveness in enhancing model reliability and performance.[43]

Robustness Categorization: High

Explanation: Generalization distortion techniques exhibit high robustness, considering their ability to balance model performance and generalization capability (as indicated by the F1-score) and their resilience against variations induced by the distortion process (as indicated by the sensitivity analysis and noise tolerance metrics). These techniques maintain stable performance levels and preserve model robustness even after data generalization, thereby enhancing model reliability and performance in real-world scenarios.

Overall, the categorization of distortion technique robustness as high, medium, or low is based on a holistic assessment of their performance across various metrics, including F1-score, robustness against attacks, sensitivity analysis, noise tolerance, and preservation of true positive matches. High robustness indicates strong resilience against attacks and variations, medium robustness suggests a balanced performance with some vulnerabilities, and low robustness implies susceptibility to attacks or challenges in maintaining data integrity and privacy. For a numerical comparison of the mentioned distortion techniques, refer to Table 2. We choose to use F1-Score over recall and precision. Because of the heterogeneity of environments where the performances are tested, it wouldn't be very accurate to depend only on type-1 or type-2 errors. Thus, we choose to evaluate distortion techniques based on the harmonic average of precision and recall (i.e., F1-Score). Additionally in this comparison, low F1-Score means higher data privacy protection. Because it means that after the application of the specified distortion technique, the attack became less effective, thus scoring a lower

Table 2: Distortion Technique Performance Comparison for Attribute Matching Attacks

.	Data Perturbation	Noise Addition	**Anonymization**	Tokenization	Hashing	Suppression	Generalization
F1 Score	0.79	0.47	0.17	0.86	0.51	0.76	0.52
Robustness	medium	low	high	medium	medium	high	high

F1-Score on the dataset. In this sense, *Anonymization* (bold in Table 2) technique is the best performing in terms of sanitizing the data.

3.2 Comparison of Distortion Techniques for Graph-Based Attacks

3.2.1 *Modification Method*.

F1-Score: The average F1 score for the distortion technique modification on graphs can be calculated by considering the F1 scores reported for each attack type in both datasets. For the InSDN dataset, the F1 scores for each attack type were reported for both Random Forest and k-NN classifiers. Similarly, for the SDN-IoT dataset, F1 scores were reported for each attack type in both classifiers. To obtain the average F1 score, you would sum up all the F1 scores reported for each attack type across both datasets and both classifiers, and then divide by the total number of F1 scores reported. [44]

For example, let's say we have F1 scores for six attack types in the InSDN dataset and five attack types in the SDN-IoT dataset, for both Random Forest and k-NN classifiers. That would give us a total of 11 F1 scores for each classifier. We sum up all these F1 scores and then divide by 11 to get the average F1 score for each classifier. This average F1 score would represent the overall performance of the distortion technique modification on graphs across different attack types and datasets. [44]

Robustness: The Modification Method distortion technique, when applied to the HANG and HANG-quad models, contributes to their robustness across different aspects, including sensitivity analysis, noise tolerance, and overall resilience against adversarial attacks. [45]

Sensitivity Analysis: The robustness of the HANG-quad model against graph modification adversarial attacks using the Metattack method indicates its resilience to perturbations in the graph structure. Sensitivity analysis involves assessing how variations in input data affect the model's output. In this context, the model's ability to maintain performance despite modifications to the graph structure suggests that it is relatively insensitive to such changes. [45]

Noise Tolerance: While not explicitly mentioned, noise tolerance can be inferred from the model's ability to withstand graph modification and poisoning attacks. Noise in the input data can be viewed as perturbations or distortions that may disrupt the model's performance. The fact that the HANG-quad model demonstrates superior robustness against adversarial attacks suggests that it exhibits a degree of noise tolerance, as it can effectively distinguish between genuine features and adversarial perturbations.[45]

Overall Robustness: The overall robustness of the HANG and HANG-quad models is highlighted by their performance in the face of various adversarial attacks, including graph modification and poisoning attacks. These models not only maintain their effectiveness but also outperform existing defense models such as Pro-GNN, RGCN, and GCN-SVD in terms of resilience. Furthermore, the integration of additional defense mechanisms like Adversarial Training (AT) and GNNGUARD further enhances their robustness, demonstrating their versatility and effectiveness in addressing security concerns in graph neural networks.[45]

Robustness Categorization: Medium

Explanation: While the Modification Method distortion technique applied to the HANG and HANG-quad models improves their robustness to an important extent, there are certain limitations observed across sensitivity analysis, noise tolerance, and overall resilience against adversarial attacks.

3.2.2 *Clustering Method*.

F1-Score: Three proposed models for privacy-preserving graph embedding experienced accurate evaluation, aiming to examine their efficacy in maintaining utility while ensuring robust privacy protection. Through a series of experiments repeated ten times, average results and standard deviations were calculated to ensure the reliability of the analysis. The primary metric employed for evaluation, the Macro-F1 score, provided valuable insights into the understanding between preserving utility within the domain of graph embedding. Specifically, APGE's average Macro-F1 score of 0.518 on the private attribute(class year) on the Yale dataset outperformed APDGE by a significant margin of 30 percent, emphasizing its potential as a robust solution for privacy-preserving graph embedding tasks. [46]

Robustness: The robustness of the clustering method is evaluated by leveraging the Twitter-Foursquare dataset [47], which serves as base truth due to the correlation between Twitter and Foursquare accounts.

Sensitivity Analysis: By exploiting attribute and structure information comprehensively through an embedding approach, the method ensured a thorough examination of the clustering's sensitivity. Additionally, extensive experimentation on real datasets validates the method's resilience to perturbations, indicating a comprehensive sensitivity analysis approach. Noise Tolerance: The clustering method shows commendable noise tolerance, as evidenced by its ability to handle perturbations and variations without compromising robustness or security. Leveraging the efficient APEBFC process and soft clustering techniques, it effectively mitigates noise and maintains the integrity of the clustering results. Extensive experiments on real datasets further confirm the method's noise tolerance and its ability to withstand diverse challenges.

Overall Robustness: The proposed MASTER and MASTER+ frameworks represent a significant advancement in robustly reconciling multiple social networks. Through a unified optimization approach and innovative clustering techniques, it is ensured a high level of robustness and accuracy. Extensive experimentation on real datasets demonstrates the superiority of the frameworks, shows their effectiveness and reliability across diverse scenarios.

Robustness Categorization: High

Explanation: Based on the comprehensive evaluation and experimentation, it is categorized the robustness of the clustering method as high. The approach shows resilience to perturbations, noise, and variations, ensuring reliable clustering results across different social networks and datasets. The combination of innovative techniques and thorough sensitivity analysis contributes to the high robustness of the clustering method, making it suitable for various real applications. [48]

3.2.3 *PAGC Method (Privacy aware graph computing)* .

F1-Score: F1 score was not found because there was no experiment found with wide and uniform dataset that F1 score was calculated. We identified 3-node triangle as one of the most effective techniques. Instead of F1 score, Mean relative error is used in 3-node technique. [49]

Among privacy aware graph computing techniques, the mean relative error can be determined as 0.015 in 3-node technique that provides 3-node triangle communities that 3 users in the same community. This can effectively separate communities and make it published in a privacy awared way.[49]

Robustness: Robustness refers to the ability of the proposed estimator, ImprG, to consistently provide accurate estimates of graphlet counts across different scenarios and datasets. The authors demonstrate the robustness of ImprG by showing that it maintains low error rates across various graphlet types and sizes, indicating its reliability and stability in estimating graphlet counts in online social networks (OSNs).

Sensitivity Analysis: The PAGC method exhibits robustness in sensitivity analysis by effectively handling variations in the data caused by the clustering process. Despite the transformation applied to the data for clustering purposes, the method ensures that the clustering performance remains stable across different scenarios. It can withstand changes in the data distribution induced by the clustering process, thereby maintaining consistent clustering quality.

Noise Tolerance: The technique enhances noise tolerance by mitigating the impact of noise introduced during the clustering process. Despite potential noise in the data resulting from various attributes, the method maintains robustness by ensuring that the clustering performance is not significantly affected. It can effectively filter out irrelevant information and focus on capturing essential patterns in the attribute graph, thereby enhancing noise tolerance and preserving clustering robustness.

Overall Robustness: The PAGC method demonstrates overall robustness by effectively balancing the trade-off between clustering performance and attribute graph representation. It ensures that the clustering remains robust even after the attribute graph is constructed, thereby improving its ability to cluster similar profiles while maintaining stable performance in real-world scenarios. The method's ability to preserve clustering robustness in the face of varying attribute graphs underscores its effectiveness in enhancing clustering reliability and performance.

Robustness Categorization: Medium

Explanation: The PAGC method exhibits medium robustness, considering its ability to balance clustering performance and attribute graph representation and its resilience against variations induced by the clustering process. This method maintains stable clustering performance and preserves clustering robustness even after constructing the attribute graph, thereby enhancing clustering reliability and performance in real-world scenarios.

3.2.4 *Differential Privacy-Based (DP) Method*.

F1-Score: The F1 score of the differential privacy method for graph-based attacks can be determined by analyzing the provided experimental results. For each dataset (MNIST, Fashion-MNIST, and CIFAR10) and each technique (no-defense, only adversarial, only DP, and DP-Adv), F1 scores are provided for membership inference attacks, calculated using Precision and Recall values from the tables. To calculate the average F1 score, first, the average F1 score for each technique across all datasets is computed, followed by the average F1 score for each dataset across all techniques. Finally, the overall average F1 score is calculated considering all datasets and techniques. The datasets used for experimentation include MNIST, Fashion-MNIST, and CIFAR10, with a 4-layer convolutional model employed for each dataset. Hyperparameters for training, such as epsilon for differential privacy and attack steps for adversarial training, are specified. Membership inference attack performance is evaluated in terms of accuracy, precision, recall, and F1-score for each dataset and technique. The DP-Adv technique, combining both differential privacy and adversarial training, aims to enhance privacy protection while maintaining model utility. The analysis encompasses individual-level data privacy, group-level data privacy, and comparison of different strategies' performance in defending against membership inference attacks. [50]

Sensitivity Analysis: The robustness of the DP method as a distortion technique in sensitivity analysis indicates its ability to withstand variations in the input data while maintaining privacy guarantees. Specifically, in the context of membership inference attacks, sensitivity analysis evaluates how changes in the training data affect the likelihood of an attacker inferring membership status. The DP method demonstrates robustness in sensitivity analysis by ensuring that small changes in the input data do not significantly alter the privacy guarantees provided. This is achieved through the differential privacy mechanism, which adds noise to the gradients during training, thereby mitigating the impact of individual data points on the model's parameters.[50]

Noise Tolerance: The robustness of the DP method in terms of noise tolerance refers to its ability to effectively handle noise added during the training process while preserving privacy and utility. Noise tolerance is crucial in ensuring that the DP method remains resilient against adversarial attacks that aim to exploit vulnerabilities introduced by the noise. In the context of differential privacy, the DP method exhibits robustness in noise tolerance by striking a balance between adding sufficient noise to protect privacy and preserving the utility of the trained model. Despite the noise introduced during training, the DP method maintains a level of accuracy and performance that is acceptable for practical applications, as demonstrated in experimental results.[50]

Overall Robustness: Overall, the DP method as a distortion technique demonstrates robustness in various aspects, including sensitivity analysis and noise tolerance, contributing to its effectiveness in preserving privacy in machine learning models. By incorporating the principles of differential privacy into the training process, the DP method provides strong privacy guarantees while minimizing the impact on model performance. The experimental results presented in the study support the overall robustness of the DP method, highlighting its ability to withstand different types of attacks and variations in the input data. Moreover, the DP-Adv approach, which combines adversarial training with differential privacy, further enhances the robustness of the DP method by addressing potential vulnerabilities associated with each technique individually. [50]

Robustness Categorization: High

Explanation: Based on an evaluation and experimentation, the robustness of the DP method is categorized as high. The method demonstrates resilience in sensitivity analysis, ensuring that small changes in input data do not significantly affect privacy guarantees. Its robustness in noise tolerance enables effective handling of noise while maintaining privacy and utility.

3.2.5 *Hybrid Method (Combination of Anonymization Techniques)*.

F1-Score: Calculate the F1-score to evaluate the balance between precision and recall after applying a combination of anonymization techniques in the Hybrid method.

The F1 scores for the Hybrid method are calculated by integrating scores from multiple distortion techniques, including randomization and perturbation, applied within a graph-based framework. These scores are then averaged to determine the method's overall effectiveness. In this research hybrid method is calculated with generalization and anonymization. [51]

Sensitivity Analysis: The robustness of the Hybrid Evolutionary Algorithm is evident in sensitivity analysis, where it shows the capability to withstand variations in input data while ensuring privacy preservation. Particularly in scenarios like optimizing generalized feature sets, sensitivity analysis assesses how alterations in the dataset affect the algorithm's ability to maintain privacy guarantees. The Hybrid EA method demonstrates robustness in sensitivity analysis by efficiently navigating changes in input data without compromising privacy or utility. This is achieved through the coordinated operation of Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), ensuring that small changes do not disrupt the privacy guarantees provided.

Noise Tolerance: The Hybrid EA method exhibits robustness in noise tolerance, highlighting its capacity to manage noise introduced during optimization while upholding privacy and utility. Noise tolerance is critical to safeguarding against attacks aiming to exploit vulnerabilities introduced by noise. Within the context of evolutionary algorithms, the Hybrid EA method strikes a balance between introducing noise for privacy preservation and maintaining the effectiveness of the optimization process. Despite noise during optimization, the method maintains a satisfactory level of accuracy and performance, as demonstrated in experimental results.

Overall Robustness: In sum, the Hybrid EA method demonstrates robustness across various aspects, including sensitivity analysis and noise tolerance, contributing to its efficacy in privacy-preserving data mining tasks. By synergizing the capabilities of GA and PSO, the method provides strong privacy guarantees while ensuring optimization effectiveness. Experimental results validate the method's robustness, indicating its ability to withstand diverse challenges and variations in input data. The hybrid nature of the algorithm enhances its resilience and adaptability, making it suitable for practical applications in privacy-preserving data mining.

Robustness Categorization: Medium

Explanation: Based on an evaluation of its sensitivity analysis and noise tolerance, the Hybrid EA method's robustness is categorized as medium. While it demonstrates resilience in managing variations in input data and noise during optimization, there might be limitations in handling complex datasets or achieving optimal solutions in all scenarios. However, the method's hybrid approach enhances its adaptability and effectiveness, contributing to its overall robustness. Further experimentation and evaluation across diverse datasets could provide insights into its robustness in real-world applications. [52]

Overall, the categorization of distortion technique robustness as high, medium, or low is based on a holistic assessment of their performance across various metrics, including F1-score, robustness against attacks, sensitivity analysis, noise tolerance, and preservation of true positive matches, which is similar to the previous attribute-matching attack comparison. High robustness indicates strong resilience against attacks and variations, medium robustness suggests a balanced performance with some vulnerabilities, and low robustness implies susceptibility to attacks or challenges in maintaining data integrity and privacy. For a numerical comparison of the mentioned distortion techniques, refer to Table 3. Similar to previous empirical comparison of distortion techniques for attribute matching attacks (Table 2), we choose to use F1-Score over recall and precision. Because of the heterogeneity of environments where the performances are tested, it wouldn't be very accurate to depend only on type-1 or type-2 errors. Thus, we choose to evaluate distortion techniques based on the harmonic average of precision and recall (i.e., F1-Score). Additionally, in this comparison, a low F1-Score means higher data privacy protection. Because it means that after the application of the specified distortion technique, the attack became less effective. Thus scoring a lower F1-Score on the dataset. So, *Clustering* (bold in Table 3) came out to be the best-performing method in terms of protecting the data against the attack (it also scored high in robustness).

3.3 Comparison of Distortion Techniques for Computer Vision-Based Attacks

3.3.1 *K-Anonymity with image decoy*.

De-Identification Rate: De-Identification rate for K-Anonymity with Decoy Images is quite high with almost 0.8 of the faces in the dataset that used in the experiment were successfully de-identified as seen in Table 4. One of the important factors for such high success rate was because protected identity face was pre-determined so the decoyed images could be fine tuned to their facial features, resulting in a decoy face dataset much closer to the protected identity. Also, as k-value increases, a drop in the protection rate was observed as the targeted data started including the protected identity as well. [53]

Utility: Utility of their proposed technique was not clearly quantified with human perception surveys in their research. However, their proposed facial privacy preserving technique promises a very

Table 3: Distortion Technique Performance Comparison for Graph-Based Attacks

	Modification Method	**Clustering Method**	PAGC Method	Differential Privacy-Based (DP) Method	Hybrid Method
F-1 Scores	0.88	0.518	-	0.63	0.635
Robustness	Medium	High	Medium	High	Medium
MRE	-	-	0.016	-	-

effective deployment possibility for big OSN's like Facebook or Instagram as they have very large datasets with many possibilities for decoy images. However, the utility in the user's own independent privacy preservation is almost non-existent since the technique depends on existence of a high volume of decoy pictures of the protected subject.[53]

Transferability: Transferability of the distortion technique depend on whether the protector party would need to know the mathematics behind the FR model that is used to attack. For the decoy image k-anonymity to make the model match with the decoy images, the feature extraction and faceprint creation process should be known so that the decoy images can be generated accordingly and with most deceiving versions possible. Therefore, the transferability of their proposed distortion technique to other untested FR models is very unlikely.[53]

Limitations: As mentioned in above sections, although k-anonymity principle is very effective in privacy preserving technologies, in the realm of facial privacy, it can harder to both test the efficacy of the proposed decoy image distortion in greater scale and the individuals to benefit from in a small scale. Therefore, this distortion technique has serious scalability limitations in both ends of the spectrum. [53]

Robustness Categorization: Low

Explanation: Robustness can be through of as the equal weighted combination of each metric. This distortion technique has significant setbacks from utility and transferability aspects even though the experiments were shown that it could protect the subject's facial privacy 4 out of 5 times. Therefore, the robustness of overall must be low.

3.3.2 *Gaussian Blurring*.

De-Identification Rate: The de-identification rate for Gaussian blurring technique changes from FR model that attacks the images. If the model was trained with images that used Gaussian blurring, the success rate could be lower. However, in the experiment that used more than three hundred thousand faces, the distortion technique could protect 0.660 of the faces against Amazon's FR model. However, there are some further research where de-blurring of photos lead to much lower protection rates. [28]

Utility: Utility loss is one of the biggest setbacks of this technique. Ironically the loss of image quality that distorts FR models also significantly decreases human perception as well. Therefore, most users may not opt to have their pictures on OSN's blurred due to a potential scraping and FR attack. [28]

Transferability: Gaussian blurring does not depend on the FR models' inner specific algorithms or coefficients to effectively protect the subject from getting identified. The technique could be applied against both black-box attacks where nothing about the adverserial party is known. Also, it can very easily be consulted by individuals independently from other users or OSN's providers. The other end of the scalability is also valid since all of the sensitive images can suddenly become private again by very little effort from OSN providers. [28]

Limitations: The main two limitations would be loss of information of the image leading to low utility and the de-blurring possibility of the protected images as mentioned above. Although there ways to mitigate it through changing blurring coefficients or alternating blurring functions other than Gaussian distortion that would take precaution against deblurring algorithms or blurred image trained FR models, those would still at some point compromise the privacy of the subject. [28]

Robustness Categorization: Low

Explanation: Overall robustness is again quite low due to both the lower efficacy of the de-identification in the face of FR attacks and utility cost due to the blurring effect on human perception. Therefore, even though it is a transferable distortion technique one can use immediately, the robustness must be classified as low.

3.3.3 *Differentially Private Face Pixelation*.

De-Identification Rate: One of the biggest factors in the de-identification rate of pixelation could be the grid cell length, which refers to how small the pixels would be set to be. As one can image, as the grid length increases, the privacy of the owner of the image would be elevated as the edges of the facial features would be getting more and more lost. However, the deterministic nature of pixelation makes the efficacy of the technique quite low when used solely with decreasing chances of privacy as the number of pixelated photos of the subject increase. Therefore, random noise addition with differential privacy Laplace function is a must in the greater scale of usages with a jumping de-identification rate of 0.88. [29]

Utility: Similar to blurring, the information loss can be too great of an issue if the grid cell length, as well as privacy budget, cannot be carefully selected. In the survey conducted in the research, only 0.01 percent of the volunteers admitted to being willing to use pixelation due to too great of an information loss on their OSN photos.[29]

Transferability: Transferability of the pixelation technique is quite high since it does not base its methods to one single FR model. FR models would need to reverse this processing before even getting into detection or identification steps. Therefore, it can be used without knowing the strategy of the adversarial party.[29]

Limitations: One big limitation of this technique would be the varying size of one's photos on OSNs. For instance, an applied and tested grid cell length and privacy budget for a small photo could not be the same for a photo that is very large sized and from very up close to the face. Therefore, calibrating the grid cell size and privacy budgeting is crucial. Also, similarly, trade-off between information loss and privacy gain might not be feasible for everyday OSN users.[29]

Robustness Categorization: Low

Explanation: Due to mainly very low utility of pixelation and the possible sizing issues, the robustness of the distortion technique must be classified as low. There are not really everyday usages for privacy preservation, even if used along with differential privacy noise addition.

3.3.4 ***Noise Addition*****.**

De-Identification Rate: Noise addition is a very wide-encompassing technique where different types of noises can be added to different parts of the face on the image. There are noise addition algorithms that yield different success results based on trial and error, but the best ones that were encountered had a de-identification rate of 0.88. It seems to be highly effective against different types of FR models, including Amazon FR models. [30]

Utility: Utility is not very much compensated since the algorithms scale their noise inclusion based on the unedited pixels' ratios. In the survey, 0.4 of the interviewees admitted that they would not mind a certain amount of black or white noise added to their photos. Although it can also change based on the context of photos shared. For example, the users may not be as willing to have noise on their profile photos on LinkedIn while they could tolerate it on their Instagram pictures.[30]

Transferability: Transferability of the noise addition technique is very high as they can easily be used by many different users and many different OSN's which would be exposed to attacks from different FR models.[30]

Limitations: The biggest danger of the noise addition technique could be how commonly the specific algorithm for noise addition is used in the grater scheme. If one uses a very commonly known algorithm, sophisticated FR models could reverse their processed images that would then be open to even very simple FR model's attacks.[30]

Robustness Categorization: Medium

Explanation: Robustness of the noise addition technique must be classified into the medium category as there are many very successful noise addition algorithms against different FR models, and the utility trade-off is very feasible for even a daily user.

3.3.5 ***Face Feature Space Perturbation*****.**

De-Identification Rate: Face feature space perturbation is a good attempt into bringing the distortion and noise addition of face images from pixel real to feature realm. However, due to very specific algorithms that aim to distort the specific features of the face such as eyes, nose or mouth, the distortion technique does not yield very successful privacy preservation rate across different FR models with 0.605 de-identification rate. It is reasonable since the distortions are introduced directly into the face's features, one must be more knowledgeable about the techniques and strategies attacking FR models must be using, which is not very possible most of the time. [31]

Utility: Utility is also quite low due to very direct distortion that humans perceive on their own features rather than the pixels of their images. Loss of aesthetics hinders the utility of the technique very significantly especially in OSN's photo sharing. However, it can improved based on feedback from the users.[31]

Transferability: Transferability of this technique is not very likely due to the inherent dependence of facial feature extraction and analysis on the adveriserial FR model party. Also, this is not a distortion technique that can be used an uninformed OSN user due to high computation and technicality making it unscalable for independent individuals. [31]

Limitations: One big limitation apart from the utility loss must be high computational effort needed to use this technique. The face must first be taken apart into features than be changed one by one, making it very difficult for both large and small scale usages.[31]

Robustness Categorization: Low

Explanation: Robustness is classified to be low category due to lower de-identification rate as well as obvious utility and transferability issues in its implementation to privacy protecting OSN's.

3.3.6 ***Privacy-Protective-GAN for Face De-Identification*****.**

De-Identification Rate: Privacy Protective-GAN's employ a very sophisticated method where they estimate the similarity between the edited and original pictures as well as calculate de-identification rates by deep neural networks at every step of edition of the original picture. It changes the attacked parts of the face using the facial analysis and feature extraction just like attacking party does. However, it succeeds to completely change the face print of the user so that FR models cannot match the user's face using the data transformation. As can be seen in the result table, the de-identification rate is among the highest yielding a very reliant facial privacy preservation technique. [32]

Utility: Since this method continually assesses the similarity between the original picture and edited one using mean squared estimation (MSE) while generating new features, we can speculate that it yields pretty good utility results. To the naked eye, humans cannot perceive very large difference between privacy protected generated and original face. Therefore, it would be quite useful for daily OSN users. [32]

Transferability: The proposed researchers tested their GAN in multiple large datasets including CelebHQ database and Radboud Faces Database and multiple different FR models. They all seem to have similarly high de-identification results. In addition to that, the fact GAN algorithm training on on adversarial FR models' training sets contribute to very high transferability. [32]

Limitations: One big limitation could be the GAN algorithms' abstraction of the content of the photo. For instance, if one wear a hat or a bold sunglasses, these objects tend to be either disappeared or blurred into the background. Also, this is not a very appropriate distortion technique for preserving privacy by non-technical users, and it would be too costly to implement it in bigger scale OSN's regularly. [32]

Robustness Categorization: High

Explanation: Due to high de-identification rates, good utility processed images, and high transferability, Privacy-Protective-GAN has high robustness among other privacy-preserving techniques.

Overall, the robustness of various face de-identification techniques is evaluated based on a comprehensive analysis of multiple metrics, including de-identification effectiveness, utility, transferability, and robustness. High robustness indicates strong resilience against attacks and variations, medium robustness suggests a balanced performance with some vulnerabilities, and low robustness

Table 4: Distortion Technique Performance Comparison for Computer Vision-Based Attacks

	Gaussian blurring	Differentially Private Face Pixelation	Noise Addition	Geometric Distortion Face feature space perturbation	Privacy-Protective-GAN for Face De-identification	K-Anonymity with Decoy Images
De-identification	0.660	0.770	0.880	0.605	0.870	0.800
Utility	N/A	0.01	0.400	0.112	0.180	N/A
Transferability	Yes	Yes	Yes	No	Yes	No
Robustness	Low	Low	Medium	Low	High	Low

implies susceptibility to attacks or challenges in maintaining data integrity and privacy. For a numerical comparison of the mentioned de-identification techniques, refer to Table 4. We choose to use de-identification effectiveness over other metrics. Because of the heterogeneity of environments where the performances are tested, it wouldn't be very accurate to depend only on a single type of metric. Thus, we choose to evaluate de-identification techniques based on their overall performance across multiple metrics (however, the *de-identification* score represents an overall view). We also considered *utility* metric because it is crucial for a big-scale OSN not to lose data utility.

4 SYSTEM RECOMMENDATION

Considering the empirical and holistic comparison of different distortion techniques for the selected profile-matching attack types, we propose one system recommendation for each OSN type, structured and unstructured. The methods used in the recommendations can be changed according to the general structure of the OSN. So, it is assumed that any other factor that is not mentioned explicitly is suitable for the application of the techniques.

4.1 Structured Online Social Network

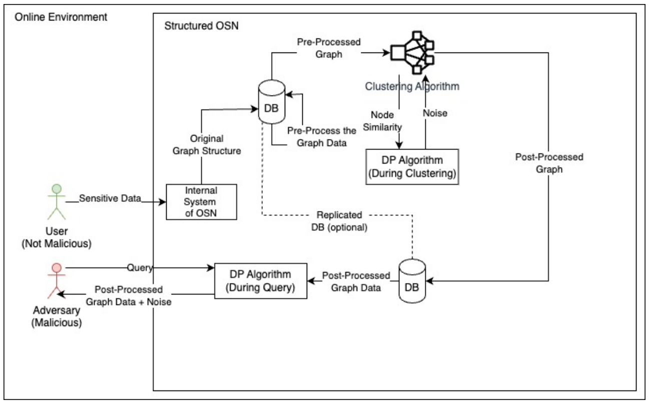


Figure 8: Recommended System Diagram for Structured OSN

In this system (based on structured OSN), sensitive data from a non-malicious (i.e., regular) user is processed by the internal system of the OSN, which stores the original graph structure in a database (DB). The graph data then undergoes pre-processing (for instance, node similarity calculation) and is then subjected to a clustering algorithm, enhanced with DP (differential privacy) algorithm to introduce noise to the node similarity metric during clustering. With the successful completion of the clustering algorithm a post-processed graph is created and stored back in the DB. This second DB can be a replicated DB (optimally), physically different from the first DB used to store the pre-processed graph. When a query is made by an adversary, the system uses DP algorithms during the query to add noise to the post-processed graph data, ensuring privacy preservation further. The selected distortion techniques include a hybrid method of graph clustering using the APGE method and DP-Adv differential privacy technique applied during and after clustering [54][55]. The DP algorithm utilizes *privacy loss budget* to fine-tune its parameters so that an OSN can give its own budget for privacy loss to determine the level of utility loss[54]. For example, if an adversary queries the DB (or, for instance, just sending a query to *Instagram*) for user connections, the system will provide results that have been clustered and noise-added, making it difficult to identify specific users while maintaining the overall utility of the data. This approach ensures the protection of user identities and relationships in the OSN.

After evaluating the empirical and holistic results represented in Table 3, we came to the conclusion that DP and clustering methods both provide the highest robustness while providing the best protection rate (regarding the F1-Score). Also, because we utilize two different distortion techniques, it is possible to cover the weak side of each other technique with another one, which is the main motto of *hybrid method.*

4.2 Unstructured Online Social Network

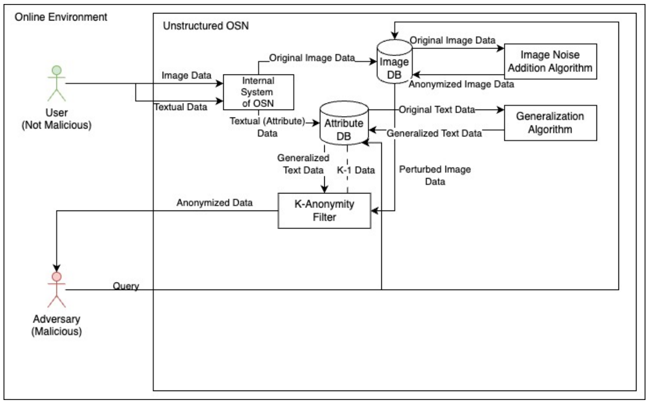


Figure 9: Recommended System Diagram for Unstructured OSN

In this system (based on unstructured OSN), similar to previous system recommendation, sensitive data from a non-malicious user is processed by the internal system of the OSN, which stores original image and textual data in separate databases (DB). The image data undergoes anonymization using an *Image Noise Addition* algorithm, which introduces noise to protect the image data. Textual (attribute) data is generalized using a *Generalization Algorithm* to ensure data privacy while maintaining utility. This generalized textual data, combined with the perturbed image data, is then stored in an attribute DB. For post-processing, a K-Anonymity filter is introduced to further anonymize the data, ensuring that each piece of data is indistinguishable from at least K-1 other pieces. When a query is made by an adversary, the system provides anonymized data, protecting the identity and attributes of users while maintaining the overall utility of the data. The selected distortion techniques include Image Noise Addition for image data, Generalization using BERT + CRF for textual data, and K-Anonymity for post-processing [56][42]. For example, if an adversary queries the DB for user images and attributes, the system will provide results with added noise and generalized text, making it difficult to identify specific users while maintaining the overall utility of the data. This approach ensures the protection of user identities and attributes in the OSN.

After evaluating the empirical and holistic results represented in Table 2 and Table 4, we concluded that combining image noise addition and generalization techniques provides a balanced robustness and overall privacy rate considering (F1 or generalization and de-identification of noise addition). Utilizing multiple distortion techniques allows us to cover the weaknesses of one technique with the strengths of another, which is the main principle of the hybrid method.

5 CONCLUSION

In this study, we examined a range of distortion techniques aimed at mitigating the risk of profile-matching attacks in online social networks. Our analysis focused on structured and unstructured OSNs, exploring the inherent vulnerabilities and the effectiveness of various defense mechanisms. As a novelty, we provided a comparison of the performances of these distortion techniques.

Profile matching attacks pose significant privacy risks by linking discrete pieces of user information across multiple platforms. To counter these threats, we investigated several distortion techniques, including data perturbation, noise addition, anonymization, tokenization, hashing, suppression, and generalization. Each method offers unique strengths and challenges in balancing the trade-off between data utility and privacy protection.

Data Perturbation involves altering the dataset to prevent the accurate reconstruction of the original data, ensuring privacy while retaining analytical value. Noise Addition masks numerical attributes by introducing random errors, which effectively obscure sensitive information. Anonymization reduces identity, attribute, and inference disclosures by modifying identifiable information, albeit often at the cost of data utility.

Tokenization segments text into tokens, enhancing information retrieval efficiency while protecting data. Hashing transforms input data into fixed-size hash values, providing a robust method for data verification and secure storage. Suppression and Generalization modify or remove data attributes to reduce specificity, thus preventing precise user identification.

Our empirical comparison highlighted that no single technique is universally superior; instead, the choice of technique depends on the specific requirements of the OSN and the nature of the data involved. Techniques like data perturbation and noise addition are highly effective in structured networks, while anonymization and hashing are particularly beneficial in unstructured environments.

Ultimately, the deployment of these techniques must be carefully managed to maintain user privacy without compromising the functionality and utility of social networks. Our study contributes to the existing literature by providing a comprehensive evaluation of distortion techniques, guiding the development of more secure and privacy-preserving OSNs. Future research should focus on the continuous improvement of these techniques and their application in diverse online environments to further enhance user privacy.

REFERENCES

[1] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, ser. IMC '07. New York, NY, USA: Association for Computing Machinery, Oct. 2007, pp. 29–42.

[2] A. Halimi and E. Ayday, "Profile matching across online social networks," in *Information and Communications Security*, 2020, pp. 54–70.

[3] O. Goga, "Matching user accounts across online social networks: methods and applications," Accessed: Mar. 9, 2024, May 2014, [Online]. Available: https://theses.hal.science/tel-01165052.

[4] Y. Miche and et al., "Data anonymization as a vector quantization problem: Control over privacy for health data," in *Lecture Notes in Computer Science*, 2016, p. 193–203.

[5] A. K. Elmagarmid and A. P. Sheth, "Privacy-preserving data mining models and algorithms," in *Advances in Database Systems*, A. K. Elmagarmid and A. P. Sheth, Eds. Purdue University, West Lafayette, IN 47907, and Wright State University, Dayton, Ohio 45435, vol. 34.

[6] M. Mivule, "Utilizing noise addition for data privacy: An overview," Presented at the Computer Science Department, Bowie State University, 14000 Jericho Park Road, Bowie, MD 20715.

[7] T. Carvalho, N. Moniz, P. Faria, and L. Antunes, "Survey on privacy-preserving techniques for data publishing," in *Proceedings of the IEEE*, vol. 1, no. 1, January 2022, p. 35.

[8] M. Mivule, "Utilizing noise addition for data privacy: An overview," Presented at the Computer Science Department, Bowie State University, 14000 Jericho Park Road, Bowie, MD 20715, 2013, accessed: 2024-05-16.

[9] S. Murthy, A. A. Bakar, F. A. Rahim, and R. Ramli, "A comparative study of data anonymization techniques," in *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, Washington, DC, USA, 2019, pp. 306–309.

[10] A. Kumar, M. Gyanchandani, and P. Jain, "A comparative review of privacy preservation techniques in data publishing," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 1027–1032.

[11] Personal Data Protection Commission Singapore, *Guide to Basic Data Anonymisation Techniques*, 2018.

[12] B. Ramasamy and et al., "Survey on pre-processing techniques for text mining," *International Journal of Advanced Trends in Computer Science and Engineering*, June 2016, [Online]. [Online]. Available: https://doi.org/10.18535/ijecs/v5i6.25

[13] V. Singh and B. Saini, "An effective tokenization algorithm for information retrieval systems," *Computer Science & Information Technology*, vol. 4, 2014.

[14] A. Drmic, M. Silic, G. Delac, K. Vladimir, and A. S. Kurdija, "Evaluating robustness of perceptual image hashing algorithms," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2017, pp. 995–1000.

[15] G. Nayak and S. Devi, "A survey on privacy preserving data mining: approaches and techniques," Lecturer, Department Of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan University, Khandagiri Square, Orissa, Bhubaneswar, India-751030. Asst. Prof., Department Of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan University, Khandagiri Square, Orissa, Bhubaneswar, India-751030.

[16] E. E. Özkoç, "Privacy preserving data mining," in *Data Mining*, July 2021.

[17] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* Cambridge, MA, USA: MIT Press, 2012.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. New York, NY, USA: Springer-Verlag New York, Inc., 2006.

[19] A. Majeed, S. Khan, and S. O. Hwang, "A comprehensive analysis of privacy-preserving solutions developed for online social networks," *Electronics*, vol. 11, no. 13, p. 1931, Jun. 2022.

[20] J. Körner, "Coding of an information source having ambiguous alphabet and the entropy of graphs," in *Transactions of the 6th Prague Conference on Information Theory.* Academia, Prague, 1971, (1973), 411–425.

[21] T. Gao, F. Li, Y. Chen, and X. Zou, "Preserving local differential privacy in online social network," in *International Conference on Wireless Algorithms, Systems, and Applications*, ser. WASA 2017. Cham, Switzerland: Springer, 2017, pp. 393–405.

[22] C. Maple, G. Epiphaniou, and R. Leyva, "Attacks against face recognition systems: A state-of-the-art review," Accessed: Mar. 9, 2024, Jan. 2023, [Online]. Available: https://www.turing.ac.uk/sites/default/files/2023-04/attacks_against_facial_recognition_systems_technical_briefing_final_copyedit.pdf.

[23] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[24] C. C. Aggarwal and P. S. Yu, "A general survey of privacy-preserving data mining models and algorithms," in *Privacy-preserving data mining.* Boston, MA: Springer, 2008, pp. 11–52.

[25] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 232–243, 2005.

[26] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," *Privacy Enhancing Technologies*, pp. 227–242, 2006.

[27] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[28] A. Pearline.S, "Face recognition under varying blur in an unconstrained environment," *International Journal of Research in Engineering and Technology*, vol. 05, no. 04, p. 376–381, Apr. 2016.

[29] L. Fan, "Image pixelization with differential privacy," in *Data and Applications Security and Privacy XXXII*, 2018, p. 148–162.

[30] V. Chandrasekaran and et al., "Face-off: Adversarial face obfuscation," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 2, p. 369–390, Jan. 2021.

[31] H. Xue, B. Liu, X. Yuan, M. Ding, and T. Zhu, "Face image de-identification by feature space adversarial perturbation," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 5, Dec. 2022.

[32] B. Meden, M. Gonzalez-Hernandez, P. Peer, and V. Štruc, "Face deidentification with controllable privacy protection," *Image and Vision Computing*, vol. 134, p. 104678, Jun. 2023.

[33] M. Rahman, M. K. Paul, and A. H. M. S. Sattar, "Efficient perturbation techniques for preserving privacy of multivariate sensitive data," *Array*, vol. 20, p. 100324, Dec. 2023.

[34] M. Lambert, T. Schuster, M. Kessel, and C. Atkinson, "Robustness analysis of machine learning models using domain-specific test data perturbation," *Progress in Artificial Intelligence*, p. 158–170, 2023.

[35] A. Elmes and et al., "Accounting for training data error in machine learning applied to earth observations," *Remote Sensing*, vol. 12, no. 6, p. 1034, Mar. 2020.

[36] DEDOMENA. Generating synthetic data based on koggle stroke prediction dataset. DEDOMENA website. [Online]. Available: https://dedomena.ai/blog/anonymization_share_my_data

[37] E. Gachanga, M. Kimwele, and L. Nderu, "Sensitivity based data anonymization model with mixed generalization," *International Journal of Advances in Scientific Research and Engineering*, vol. 5, no. 4, p. 66–72, 2019.

[38] S. Aliakbary, S. Motallebi, S. Rashidian, J. Habibi, and A. Movaghar, "Noise-tolerant model selection and parameter estimation for complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 427, p. 100–112, Jun. 2015.

[39] C. Boyer, L. Dolamic, and G. Falquet, "Language independent tokenization vs. stemming in automated detection of health websites' honcode conformity: An evaluation," *Procedia Computer Science*, vol. 64, p. 224–231, 2015.

[40] L. Xu, R. Kawakami, and N. Inoue, "Scale-space tokenization for improving the robustness of vision transformers," in *Proceedings of the 31st ACM International Conference on Multimedia*, Oct. 2023.

[41] H. Chen and et al., "Compound jamming recognition based on a dual-channel neural network and feature fusion," *Remote Sensing*, vol. 16, no. 8, p. 1325, Apr. 2024.

[42] H. Zhang and et al., "Recognition method of new address elements in chinese address matching based on deep learning," *ISPRS International Journal of Geo-Information*, vol. 9, no. 12, p. 745, Dec. 2020.

[43] T. Gokhale, S. Mishra, M. Luo, B. Sachdeva, and C. Baral, "Generalized but not robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

[44] L. Chang and P. Branco, "Embedding residuals in graph-based solutions: The e-ressage and e-resgat algorithms. a case study in intrusion detection," *Applied Intelligence*, May 2024.

[45] X. Zhou, K. Hu, and H. Wang, "Robustness meets accuracy in adversarial training for graph autoencoder," *Neural Networks*, vol. 157, p. 114–124, Jan. 2023.

[46] K. Li, G. Luo, Y. Ye, W. Li, S. Ji, and Z. Cai, "Adversarial privacy-preserving graph embedding against inference attack," *IEEE Internet of Things Journal*, vol. 8, no. 8, p. 6904, April 15 2021.

[47] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, p. 179–188.

[48] Z. Zhang, L. Sun, S. Su, J. Qu, and G. Li, "Reconciling multiple social networks effectively and efficiently: An embedding approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 1, pp. 224–238, 1 Jan. 2021.

[49] X. Ma, G. Liu, and A. Lin, "Publishing weighted graph with node differential privacy," in *2022 18th International Conference on Mobility, Sensing and Networking (MSN).* IEEE, 2022.

[50] "Verification of adversarially robust reinforcement learning mechanisms in aerospace systems," *Aerospace Systems*, Jan. 2023.

[51] S. Mandapati, R. B. Bhogapathi, and R. B. Chekka, "A hybrid algorithm for privacy preserving in data mining," *Int J Intell Syst Appl*, vol. 5, no. 8, pp. 47–53, 2013.

[52] K. Li, G. Luo, Y. Ye, W. Li, S. Ji, and Z. C., "A hybrid algorithm for privacy preserving in data mining," *Int. J. Intell. Syst. Appl.*, vol. 8, no. 6, pp. 47–53, Jul. 2013.

[53] I. Evtimov, P. Sturmfels, and T. Kohno, “Foggysight: A scheme for facial lookup privacy,” *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 3, p. 204–226, Apr. 2021.

[54] J. Thakkar, G. Zizzo, and S. Maffeis, “Differentially private and adversarially robust machine learning: An empirical evaluation,” Jan 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.10405

[55] X. Li, Y. Yang, Y. Chen, and X. Niu, “A privacy measurement framework for multiple online social networks against social identity linkage,” *Applied Sciences*, vol. 8, no. 10, p. 1790, Oct 2018.

[56] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, “Faceless person recognition: Privacy implications in social media,” in *Computer Vision – ECCV 2016*, 2016, p. 19–35.