Data Anonymization in Social Networks

Ouafae Baida, Mariam Ramdi, Oumaima Louzar and Abdelouahid Lyhyaoui

January 4, 2020

Data Anonymization in Social Networks

State of the Art, Exposure of Shortcomings and Discussion of New Innovations

Baida Ouafae*
LTI Lab, ENSA of Tangier
Abdelmalek Essaâdi University
Tangier, Morocco
wafaebaida@gmail.com

Ramdi Mariam*
LTI Lab, ENSA of Tangier
Abdelmalek Essaâdi University
Tangier, Morocco
mariam-ramdi@outlook.com

Louzar Oumaima
LTI Lab, ENSA of Tangier
Abdelmalek Essaâdi University
Tangier, Morocco
Louzar.oumy@gmail.com

Lyhyaoui Abdelouahid
LTI Lab, ENSA of Tangier
Abdelmalek Essaâdi University
Tangier, Morocco
lyhyaoui@gmail.com

***Abstract*—Privacy is a concern of social network users. Social networks are a source of valuable data for scientific or commercial analysis. Therefore, anonymizing social network data before releasing it becomes an important issue. The nodes in the network represent the individuals and the links among them denote their relationships. Nevertheless, publishing a social graph directly by simply removing the names of people who contributed to this graph raises important privacy issues. In particular, some inference attacks on the published graph can lead to de-anonymizing certain nodes, learning the existence of a social relation between two nodes or even using the structure of the graph itself to deduce the value of certain sensitive attributes. In this paper, we present a brief yet systematic review of the existing anonymization techniques for privacy preserving publishing of social network data. We identify the challenges in privacy preserving publishing of social network data comparing to the extensively studied relational case. We survey the existing anonymization methods for privacy preservation in three categories: graph modification approaches, generalization approaches and differential privacy methods.**

***Keywords*—Privacy, Social networks, Anonymizing, Publishing, Attacks, Graph, De-anonymizing, Sensitive attributes, Data utility, Graph modification, Generalization, Differential Privacy.**

I. Introduction

Recently, social network platforms have particularly attracted users thanks to their easy access and advanced features. There are various social networking sites such as Facebook, Google Plus and LinkedIn that allow users to create their profile and maintain their connections. People use these platforms to share their thoughts, images, videos, and search for old friends and create new ones. They also use them to subscribe to their favorite community or join a group.

According to Facebook statistics, there would be more than 2.4 billion monthly users and 1.56 billion daily users [1]. As social network data becomes more easily accessible and collected, many web access providers publish this data for research purposes. The analysis of social networks is used in modern sociology, geography, economics, and information science as well as in various fields. However, publicizing the original data of social networks raises issues of confidentiality. The adversary can search for documented threats such as identity theft, digital harassment and personalized spam. Since this user information is publicly available, it can also be used to form predictive models that can derive private information from the user and also predict its behavior. Many works have been proposed for the publication of data on social networks preserving confidentiality.

Three types of users are involved to make the data accessible to the public, namely. Owners of data that share information on social networks, service providers responsible for collecting and managing social network data, and third parties interested in using the data. Online media service providers have different motivations, they can hope that data mining will provide additional functionality to their users or produce useful results that they can share with others. While third parties are interested in user data for marketing, advertising, or data collection and resale, some of them have malicious intent.

Even though the published social network dataset is useful for a specific search inquiry, the combination of several datasets and some basic information can infer the privacy of the users. The field of privacy-preserving data publishing research investigates how to publish the data in a manner that preserves the privacy of the user whose records are being published, while maintaining the published dataset rich enough to allow for the exploration of data. Most data disclosure research focuses on the protection of individuals against the dissemination of sensitive attributes that could be troublesome or harmful, if any. Traditional examples of sensitive data include medical records and criminal records. However, the goal of the social network is somewhat different. The goal is to prevent a user from being identified by his published data and to protect his sensitive data.

Research efforts devoted to the protection of privacy have given rise to several methods and variants of models. For example, Fung et al. [2] lists no less than fifteen models. Here are some examples to counter the different types of attacks, and among them the most referenced models in the literature, namely k-anonymity [3], l-diversity [4], t-proximity [5] and δ-presence [6]. Although privacy respect for relational databases publication has been widely studied and several anonymization

* These authors contributed equally to this work

techniques exist with varying degrees of reliability and contexts of applicability such as Attribute Suppression [7], Generalization [8], Data swapping [9], Random noise [10], Character Masking [11] and Pseudonymization [12]. In this paper, we will concentrate about anonymization techniques of social network data.

II. Data anonymization and social networks

Many naive users may do not know that the information they provide online is stored in massive data repositories and may be used for various purposes. Researchers have pointed out the privacy implications of massive data gathering, and a lot of effort has been made to protect the data from unauthorized disclosure. However, most of this work has been on micro data (data stored as one relational table, where each row represents an individual entity) and models such ask-anonymity and l-diversity have been developed for it. But these models cannot be simply applied to social network data. Anonymization of social network data is a much more challenging task than that of micro data. Firstly, in relational (micro data) databases, attacks come from identifying individuals from quasi-identifiers. But in social networks, information such as neighborhood graphs can be used to identify individuals. Secondly, tuples can be anonymized in relational data without affecting other tuples. But in social networks, adding edges or vertices affects the neighborhoods of other vertices in the graph as well.

Technological advances have made social network data collection very easy. However, agencies and researches that collect such data often face with two undesirables problems. They can publish data for others to analyse, but this will create severe privacy threats, or they can withhold data because of those privacy concerns, but this will make further analysis impossible. Therefore, the goal is to enable the useful analysis of social network data while protecting the privacy of individuals. The published data may contain some sensitive data of individuals in the social network, which must not be disclosed. For this, social network data must be anonymized before it is published. This anonymization should offer protection against potential re-identification attacks. Even then, graph structure and background knowledge combine to threaten privacy in many new ways [13].

A. Social Network Representation

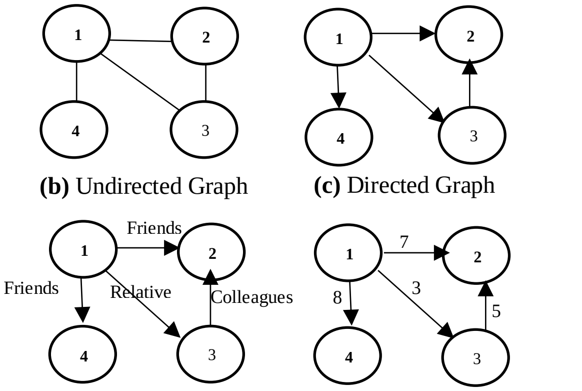
Social network analysts use two methods to represent social networks: graphs and matrices. The first method, graphs, consists of points (or nodes) to represent actors and lines (or edges) to represent ties or relations. These are called sociograms and are a very useful way of representing information about social networks. However, it becomes hard to see patterns when there are many actors and/or many kinds of relations. The other method used to represent social networks is matrices. Matrices allow the application of mathematical and computer tools to summarize and find patterns. The most common form of matrix for social network analysis is the adjacency matrix. The graph with 'n' actors is represented as an adjacency matrix of size n×n. A relationship between ith and jth node is represented by the value in the cell i, j. Graphs can Handle large social networks, provide a rich vocabulary to easily model social networks (labels, values, weights, etc.) as well as mathematical operations that can be used to quantify structural properties and prove graph-based theorems. But Signed and valued graphs have to be used to represent valued relations. Matrices are efficient just for small networks and Easy to denotes ties between a set of actors (a matrix for each relationship) but Not a best choice for large social networks and Difficult to use when network data contain information on attributes.

Due to the large size of social networks, matrices are not the most appropriate way to represent these networks. In this paper, we model a social network as a graph G = (V, E) enriched by a set of attributes A, where V represents the set of vertices as each vertex corresponds to an individual, and E a set of edges as each edge represents a social relation (friendship, common interests, sexual relations, financial exchanges, enmity, etc.) between two individuals. The set of attributes A is such that for every vertex in V we can find attributes such as name, telephone number, age, etc., and for each edge in E, we can characterize it by an attribute such as the type of relationship.

To represent different forms of data and to model the structural properties of social networks, graphs can have their edges and nodes labeled or unlabeled, directed or undirected, weighted or unweighted as explained in what follows.

V	**1**	**2**	**3**	**4**
1	0	1	1	1
2	1	0	1	0
3	1	1	0	0
4	1	0	0	0

(a) Matrix sample case to represent social network actors and their relationships



(b) Undirected Graph **(c)** Directed Graph

(d) Labeled Directed Graph **(e)** Weighted Directed Graph

Fig. 1. A social network representation using a matrix, (a), an undirected graph (b), a directed graph (c), a labeled graph (d), and a weighted graph (e) with n = 4 nodes and m = 4 links.

B. Purpose of anonymization and utility

The anonymization process, regardless of the techniques used, reduces the original information in the dataset. And generally, as the anonymization increases, the utility of the dataset decreases. the degree of arbitration between an acceptable (or expected) utility must be determined and the risk of re-identification reduced if the data subject is identified from data that should be anonymized, it should be noted that utility should not be measured at the dataset level but is usually different for different attributes. One extreme is that a specific attribute is the main element of interest and no generalization [14] / anonymization technique should be applied, while the other extreme might be that a certain attribute is unusable for the purpose intended and may be deleted entirely without affecting the usefulness of the data to the recipient.

Another important consideration in terms of usefulness is whether this poses an additional risk if the recipient knows what anonymization technique and degree of granularity were applied, firstly, it could help the analyst to better understand or interpret the results, but on the other hand, it may contain clues that may increase the risk of re-identification. So far, two types of utility as follows have been considered.

- General graph properties: One of the most important applications of social network data is analyzing graph properties. For example, researchers may be interested in the distribution of vertex degrees in a network. Some other graph properties that are often used include diameter and clustering co-efficient of networks. Some of them are addressed in [15], [16], [17], [18], [19], [20], [21].
- Aggregate network queries: An aggregate network query [22], [23], [24] computes the aggregate on some paths or subgraphs satisfying some given conditions. As an example, suppose a user is interested in the average distance from a medical doctor vertex to a teacher vertex in a social network. For each doctor vertex, we can find the nearest neighbor vertex that is a teacher. Then, the aggregate network query returns the average of the distance between a doctor vertex to its nearest teacher neighbor. Aggregate network queries are useful in many applications, such as customer relationship management.

C. *Challenges in Anonymizing Social Networks Data*

Privacy preservation for social network data is much more challenging and complex than relational data. Tuples in a relational table are independent of each other.

Models such as k-anonymity and l-diversity have been developed for privacy preservation in relational data. But these cannot be applied to social network data straightforwardly. Anonymization of social network data is a much more challenging task than anonymizing relational data. Firstly, in relational databases, attacks come from identifying individuals from quasi-identifiers. But in social networks, information such as neighborhood graphs can be used to identify individuals. Secondly, tuples can be anonymized in relational data without affecting other tuples. But in social networks, adding edges or vertices affects the neighborhoods of other vertices in the graph as well [25].

D. *Adversary Knowledge*

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

The adversary uses a variety of background knowledge to encroach on the privacy of social networks. The adversarial background knowledge plays an important role in understanding the type of the attacks as well as the various protection methods. The background knowledge has referred as information of network data that an adversary imposes a privacy attacks on the published social network data. The adversary can obtain this type of information by crawling or by exploring the overlapping membership of several social networking sites or by stealing the web browsing history, which can be used to re-identify a particular person in the published social network data.

The personal attributes represent the non-structural information that describes social network users (e.g., name, address, age, salary, marriage status, etc.). These attributes are assigned to the vertex or edge. Some of the personal attributes such as social security number act as a unique identifier. The network user removes these types of attributes before publishing the data. Other personal attributes such as name and address act as quasi-identifiers. Quasi-identifiers may not be sensitive, but an adversary can combine them with other information (e.g., Auxiliary information) to mount sensitive information disclosure attack on the published social network data. The structural attributes represent the graph information like degree, neighborhood and some other properties which can help an adversary to accomplish privacy attacks on anonymous graphs. The degree of a vertex $\boldsymbol{v}$ is the number of edges incident to that vertex and is represented as $\mathbf{deg}(\boldsymbol{v}) = |\{\boldsymbol{u}| (\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E}\}|$ of a graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$. The number of neighbors of a vertex $\boldsymbol{v}$ is the set of vertices adjacent to the vertex $\boldsymbol{v}$ and it is represented as $(\boldsymbol{v}) = |\{\boldsymbol{u}|\boldsymbol{evu} \in \boldsymbol{E}\}|$. These metrics are simple where the adversary can easily obtain and uses as a background knowledge to perpetrate privacy attacks. The auxiliary information (also referred to as external knowledge) is the information that an adversary is gathered from other sources such as another social network graph which has overlapping users with the published social network graph and group membership of users. An auxiliary social network graph which has overlapping users with the published social network for de-anonymization is also used in [26], [27], [28]. It has been shown that the auxiliary information can be used for a substantial re-identification attack even if it is very noisy. The adversary also uses a subgraph structure as a background knowledge to breach the privacy from anonymous graphs. For a given **Graph** $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, a subgraph is $\boldsymbol{H} = (\boldsymbol{V}', \boldsymbol{E}')$ where $\boldsymbol{V}' \subseteq \boldsymbol{V} and\ \boldsymbol{E}' \subseteq \boldsymbol{E}$. It contains no vertices or edges that are not in the original network. An embedded subgraph includes subgraphs and special edges within the target social network. In summary, the adversary can use a wide variety of background knowledge to mount an attack on published social network data. It is not possible to model all types of the adversary knowledges and the type of the published graph determines the use of the adversary knowledge [29].

E. *Graph anonymization techniques*

From a high-level view, the privacy preservation methods can be classified as Graph Modification Methods, Generalization or Clustering Methods and Differential Privacy Models.

1) Graph Modification Methods: Graph modification approaches anonymize a graph by modifying (that is, inserting and/or deleting) edges and vertices in a graph. The modification can be conducted in three ways and correspondingly there are three sub-categories of the methods. The optimizations approaches try to make up an optimal configuration and modify the graph accordingly. The randomized graph modification approaches conduct perturbation. Last, the greedy graph modification approaches greedily introduce modification to meet the privacy preservation requirement and optimize the data utility objectives [30].

a) Randomization Techniques: In this anonymization, the original graph is modified randomly by adding noise either by adding, deleting, switching edges or vertices and their attributes. Randomization techniques protect against re-identification in a probabilistic manner. Generally, graph randomization techniques can be applied to remove some true edges and/or adding some false edges. One of the strategies is Rand add/del method in which randomly adds one edge followed by deleting another edge which preserves the number of edges in the original graph. Secondly, Rand Switch method in which selects a pair of existing edges (vi, vj) and (vm, vn) randomly and switch the edges to (vi, vn) and (vm, vj) where (vi, vn) and (vm, vj) edges do not exist in the original graph. The Rand Switch method preserves the number of edges and degree of each vertex. There are different randomization approaches proposed for privacy preservation in social networks [31], [32], [33]. Ying and Wu proposed Spectrum Add/Del and Spectrum Switch randomization methods specifically designed to preserve the spectral characteristics of the original graph [34]. In addition, the authors also developed a variation of the Random perturbation method, called Blockwise Random Add/Delete (Rand Add/Del-B) method in which the algorithm divides graph into blocks according to the degree sequence and implements modifications by adding or removing edges on the high risk of re-identification, not at random over the entire set of vertices. Bonchi et al. [35] Proposed a new information theoretic perspective on the level of anonymity obtained by randomization methods. They made an essential distinction between image and pre-image anonymity and used entropy quantification to measure the level of anonymity provided by the perturbed graph.

b) K-anonymization Techniques: Most of the graph modification approach uses a k-anonymization method in which the models provide anonymity by adding or deleting edges or vertices of a graph to meet some certain constant value. There are different k-anonymity based methods that primarily differ in the adversary background knowledge have been developed to mitigate the vertex re-identification.

c) Degree Based Anonymization Techniques: Generally, one of the main graph properties is the degree of a vertex. In degree-based anonymization approaches the adversary uses the degree of a vertex as background knowledge to identify the particular vertex in the graph. For example, assume that an adversary knows that a target vertex has 4 adjacent vertices in the network. In the naïve anonymized graph, if there is only one vertex has the degree 4 then the adversary can re-identify the targeted vertex

*d) K-degree anonymity**:*** A Graph $G = (V, E)$ is said to be a k-degree anonymous if for every vertex $v \in V$ in graph G there are at least $k - 1$ other vertices have the same degree of graph G. The k-degree anonymization problem can be achieved by transforming the original graph G into k-anonymous graph G' with only adding edges or adding fake vertices or both. In these cases, the main optimization is to minimize the number of newly added edges and vertices to preserve the much of the characteristics of the original graph. Lu et al. [36] proposed a greedy algorithm, called Fast k-degree anonymization algorithm that anonymizes the original graph by interleaving the anonymization of the degree sequence with the construction of anonymized graph. Chester et al. [37] [38] proposed k-degree anonymization by adding only fake vertices rather than edge set. The algorithm creates links between fake vertices and original vertices or between fake vertices in order to achieve the k-anonymity. The fake vertices also must be k-anonymous.

e) Neighborhood Based Anonymization Techniques: In this case the adversary uses the background knowledge of the immediate neighbors of a vertex to disclose the identity of individuals. There are several approaches have been developed for neighborhoods-based attacks of social network data publishing. The neighborhood vertex $v \in V$ of a graph G is a subgraph of the neighbors of vertex v of the original graph [22] [23].

f) K-neighborhood anonymity: A graph $G = (V, E)$ is k-anonymous if for every vertex $v \in V$ is k-neighborhood anonymous in G if there are at least $k - 1$ other vertices in the graph such that $N(v1\ )$, $N(v2\ )$, … $N(vk{-}1)$ are isomorphic where $N(vi)$ is a neighborhood subgraph of vertex $\boldsymbol{vi}$**.** Sun et al. [39] identified a mutual friend attack problem where the adversary knows the number of common neighbors between two connected vertices. They proposed an edge anonymization k-NMF algorithm in which they ensure for each connected edge $e \in E$ there exist at least $k - 1$ other edges that share the same number of common neighbors of e in the graph.

g) Subgraph Based Anonymization Techniques: In this case the adversary uses the subgraph as a background knowledge in which to identify a targeted individual in the original graph. In this the adversary model the knowledge as a query Q that result to a subgraph of the graph G and disclose the vertex identity without the prior structural knowledge of the graph. This can be formalized by the notion of graph automorphism.

h) K-automorphism: A Graph $G'\ (V', E')$ is said to be k-automorphic such that for each vertex v there exist at least $k - 1$ automorphic functions $\{f1, f2, \dots fk{-}1\}$ of G' and $fi\ (v) \neq fj(u)$ where $v \neq u$ and $i \neq j$. Zou et al. [40] proposed the k-automorphism to solve the subgraph-based privacy attacks. The anonymization model preserves the privacy by providing at least k- structurally identical subgraphs in the published graph. This approach constructs a graph in which each vertex $v \in V$ is automorphic to at least $v1, v2, \dots vk{-}1$ other vertex in the graph. This can be achieved by the process of alignment of subgraphs and addition of edges in the graph. This approach partitions the original graph into a set of unique subgraphs such that each subgraph contains at least k-subgraphs and no subgraphs share a vertex. k-automorphism model able to guarantee privacy under any structural attack. The k-automorphism ensures that the anonymized graph at least $k-1$ automorphism functions such that each function map every vertex to a different other vertex. The information loss is measured as an anonymization cost as defined below: $cost$

$$\boldsymbol{(G, G') = (E(G) \cup E\ (G\ ') - E(G) \cap E\ (G\ '))}\ (1)$$

Where $E(G)$ indicates number of edges in graph G. The lower cost is an indication of fewer changes to the original graph G.

2) Generalization Techniques: These anonymization techniques are based on the idea of clustering vertices and edges into groups and then form a super-vertex. The inconvenience of the clustering-based methods is that the graph may be shrunk after anonymization and local structures will be difficult to analyze. There are three main classes of clustering-based approaches.

a) Degree *Vertex clustering methods:* Vertex clustering methods consist in delivering an anonymized graph which is a generalized graph of the original one, with a super node instead of an original group of nodes. In Hay et al. [15] the nodes of the graph are partitioned into disjoint sets. These nodes are considered as super nodes since they are nodes of a generalized graph. The partitioning of nodes is performed such that the resulting generalized graph maximizes utility and preserves privacy.

b) Edge clustering methods: Edge clustering methods consist in delivering a representation of the original graph wherein the relational information exists between clusters of vertices. This method consists in leaving the set of edges intact. The edges will only exist between the clusters of vertices [20].

c) Vertex and Edge Clustering Methods: Vertex and edge clustering methods consist in partitioning original graph into clusters then combining nodes into a generalized node and edges between clusters into a single edge. In Campan and Truta [42] data of the graph to be anonymized is clustered. For each cluster, the corresponding subgraph is extracted and the nodes of the subgraph are collapsed into a single node. The information about the number of nodes in the cluster is attached to this generalized node as well as the number of edges in the original cluster. Then the inter-cluster edges will be collapsed into a single edge and the structural information released will limit to the total number of edges collapsed into a single edge between the two clusters.

3) Differential Privacy Models: Differential Privacy is one of the standard privacy models which is different from the previously described models. All the privacy preservation methods discussed so far will be based on the adversary background knowledge, but the differential privacy model does not depend on background knowledge. Differential privacy relies on some query and result perturbation in order to provide privacy guarantee. This can be achieved in differential privacy is by adding some random noise to the query output. This is realized by using the methods such as Laplace distribution and the normal distribution with variance depending on ϵ and the query's sensitivity. In social network data publishing, the main goal of differential privacy is to guarantee that an adversary in possession of the published results will not be able to determine that a target vertex appears in graph G or a vertex $i \in V$ and $j \in V$ are friends in the original graph $G = (V, E)$. There are various algorithms have been developed to release statistics about social network data. They are categorized into node privacy and edge privacy methods.

a) Node Differential Privacy: A privatized query Q satisfies node-privacy if it satisfies differential privacy for all pairs of graphs $G1 = (V1, E1)$, *$G2 = (V2, E2)$ where $V2 = V1 - x$ and $E2 = E1 - \{(v1, v2) \mid v1 = x \vee v2 = x\}$ for some $x \in V1$.* In node privacy, If the social network graph G can be obtained from another graph G' or vice versa by adding or deleting a node and all edges corresponding that node then the graphs are said to be node neighbors to each other. This privacy guarantee completely protects all individuals. Node differential privacy provides protection to the nodes as well as to their adjacent edges. There are different approaches have been proposed to achieve the node differential privacy. Hay et al. [43] introduced the notion of differential node privacy and draw attention to some of the difficulties in attaining it. Smith and Raskhodnikova [44] discuss a node differential privacy algorithm for releasing an approximation to the degree distribution of a graph and also discussed the approaches for analyzing the accuracy of proposed algorithms on real networks

1) Edge Differential Privacy: A privatized query Q satisfies edge-privacy if it satisfies differential privacy for all pairs of graphs $G1 = (V1, E1)$, $G2 = (V2, E2)$ where $V1 = V2$ and $E2 = E1 - x$ where $|Ex\ | = k$. In edge privacy, G and G' are said to be edge neighbors if G' can be obtained from the G if k arbitrary edges are removed or added from G. Therefore, the edge differential privacy ensures that the adversary will not be able to disclose the presence or absence of a particular edge in the graph. Nissim et al. [45] considers differential edge privacy in the case of estimating the cost of minimum spanning tree and the number of triangles in a graph and they also discussed algorithms for computing the smooth sensitivity of statistics in a variety of domains. Rastogi et al. focused on differential edge privacy for the case of general subgraph counts release against Bayesian adversary

III. Related work and discussion

Privacy in online social networks is a recent research area that is still under development. The main objectives of an anonymization process are: (1) to preserve the privacy of users or individuals who appear in a dataset, hindering the re-identification processes, and (2) to preserve data utility on anonymized data, that is minimizing information loss. In this section, we are discussing some previous works on preserving privacy of published social networks. As it already mentioned in Section E the methods for anonymizing social networks can be broadly classified into three categories: Graph modification techniques, Generalization approaches and Differentially private approaches [35].

A. *Graph anonymization approaches*

There are many methods to anonymize the users' identities in the social network. k-anonymity introduced by Sweeney [3] is one of the firstly proposed methods to prevent from identity disclosure. It is used in order to anonymize information to prevent structural attacks against identifying the degrees of nodes in the social network graph. In order to adapt this method to social network graphs, a k-degree anonymity method was proposed [17]. in this method, the k-anonymity notion was used on the vertices. that is, for each vertex in the graph, there are at least k-1 other vertices with the same degree. To apply k-degree, different methods have been proposed which are innovative. Xue S Lu et al. [46] proposed one of these methods and consider its speed and scalability using several heuristic methods, it anonymizes a graph by simultaneously adding edges and anonymizing its degree sequence in groups of vertices. An advancement of this algorithm was introduced by Hartung et al [47]. Bin Zhou et al. [22] used a greedy method to obscure vertex labels. This method adds fake edges to the nodes of the graph. The number of vertices of the graph remains unchanged.

B. *Generalization approaches*

Hay et al. [48] proposed anonymizing a graph by generalizing it partitioning the nodes and summarizing the graph at the partition level. They how that a wide range of important graph analyses can be performed accurately on a generalized graph while protecting against re-identification risk.

C. *Differential privacy approaches*

Differential privacy methods are based on introducing random noise in the original data, that is randomly add/delete edges from the graph (unchanged number of edges) Or Randomly Switch edges between pairs of nodes (Unchanged degree of all nodes and number of edges). There are several works on graph randomization in literature, such as [43], [16], [18], [49], [50].

D. *Recent Research*

An extensive study about specific privacy-preserving methods and their particularities is beyond the scope of this work. However, some interesting surveys were made and can help to extend this brief summary.

In [51], Feng Li et al. Designed a comprehensive differentially private graph model that combines the dK-1, dK-2, and dK-3 series together, the goal was to preserve the structural utility as much as possible while satisfying-differential privacy by adding sufficient noise to the dK model and reconstruct a graph G based on the perturbed dK series. Peng Liu et al [52]. also used noise addition technique, they proposed an algorithm that locally adds noise to the possibility of the presence of edges in the community. They compare their method to the global differential privacy technique, they success in increasing the utility of data because of adding less noise to the data. In [53], R Kaur et al. used the machine learning classification technique on imbalanced dataset in two steps. The first step desire to get a predefined class label, and the second, was to use the classifier constructed in the previous step for the classification and the prediction of new instances based on the patterns examined in the training set. S. M Mazinani et al. [54] proposed a new algorithm for adding noise nodes to achieve k-degree and making un-unique information for social networks servers at the time of generating the social network with least changes in main graph attributes. In [55] Alex X. Liu et al. Proposed a random matrix approach that achieves both space efficiency and utility preservation. They obtain a good percent of accuracy at the 3 levels of their algorithm by adding a small amount of noise as a first step. Second, proving that the amount of added noise is small. Finally, validating their random matrix approach on three different applications: node clustering, node ranking, and node classification. J Casas-Roma et al. [56]. proposed a greedy algorithm that is driven by two criterion measures: minimization of generalization information loss and minimization of structural information loss. It is a three-step based approach. They start with bucketization, by choosing predefined attribute variables, and nodes which have the same values, then build clusters and calculate their average information loss score tuple. Finally, create the super-nodes and super-edges according to the partitions created by the best clustering distribution in the previous step. In [57], T Gao et al. proposed the bottom sketch algorithm to prevent second-round ADS (All-Distance Sketch) attack on unweighted graphs. By generating the ADS sets and graphs in the first step, then Adding/deleting edges in the second step and strike the balance between utility and privacy. T Gao et al [58]. Defined the notion of group-based local differential privacy on undirected graphs, by resolving the network into 1-neighborhood graphs and applying HRG-based methods (Hierarchical Random Graph), their scheme preserves differential privacy and reduces the noise scale on the local graphs by adding and deleting enough edges until satisfying their privacy demand.

Based on the above analysis, generalization methods need less computational time. So, it can be employed on large graphs. Regardless of increasing the privacy in this method, the utility of the anonymous graph decreases a lot, it means a high level of information loss which needs to be mitigated accordingly. Differential privacy also has caused much loss of network structure information contrariwise but it is one of the most remarkable techniques, since it could theoretically achieve a strong privacy guarantee, it was found that the modification method with constraints gave the best trade-off for information loss and risk of disclosure.

The table above represents the different research done on the of social networks anonymization using the different methods mentioned in Section E. The methods carried out between 2007 and 2014 are very important according to discoveries and number of citations, of which those which were carried out between 2018 and 2019 are based on them or on their proposed techniques. But through a large number of analysis and experiments, the results show that the existing anonymous technologies still can't resist the current graph de-anonymization attacks, more efforts are needed to ameliorate this resisting problem by preserving data utility which is minimizing information loss by keeping structural properties not much changed or those properties can be reconstructed from the anonymized graph and the privacy of users or individuals on the resulting graph.

TABLE I. COMPARISONS OF THE ANONYMIZATION TECHNIQUES.

Model	Year/ Citations	Anonymization Method	Anonymization technique	Data	Input	Output
[16]	2007/ 401	Differential privacy	Random perturbation	Hep-Th, Enron, Net- trace, Net-common	(1): Original graph	(1): Randomly perturbed graph
[17]	2008/ 805	Graph modification	k-anonymity: anonymizing vertices	Random graphs, Small-world graphs, Scale-free graphs, Prefuse graph, Enron graph, Powergrid graph, Co-authors graph	(1): Original graph	(1): K-degree anonymous graph. Dynamic programming.
[22]	2008/ 775	Graph modification	k-anonymity: adds fake edges	High Energy Physics	(1): Original graph	(1): K-neighbourhood anonymous graph.
[15]	2008/ 594	Generalization	Partitioning the nodes and summarizing the graph at the partition level	Hep-Th, Enron, Net-trace	(1): Original graph	(1): generalized graph
[34]	2008/ 334	Differential privacy	Spectrum randomization	US politics book data	(1): Original graph	(1): Spectrum preserving randomized graph
[43]	2009/ 245	Differential privacy	Graph degree distribution	Flickr, LiveJournal, Orkut, and YouTube	(1): Original graph	(1): the perturbed graph
[49]	2009/ 72	Differential privacy	Edge randomization	Polbooks, Polblogs, Enron	(1): Original graph	(1): the perturbed graph
[59]	2010/ 104	Generalization	Nodes Generalization	HepTh, Enron NetTrace, HOT, Power-Law, Tree, Mesh	(1): Original graph, minimum supernode size	(1): the generalized graph
[36]	2012/ 48	Graph modification	k-degree: adding edges and anonymizing its degree sequence	Email-Enron	(1): v: sorted vertices by degree, i: an index, k: the value of anonymity	(1): k-degree anonymous graph
[50]	2014/ 12	Differential privacy	Edge set modification according to edge's relevance	Zachary's Karate Club, US politics book data (Polbooks), URV email	(1): Original graph	(1): the perturbed graph
[60]	2015/ 70	Graph modification	k-degree anonymity (vertex and edge modification)	ca-HepTh, ca-CondMat, email-Enron, ca-AstroPh, ca-GrQc	(1): original degree sequence d, anonymization level k	(1): Anonymized Graph
[61]	2016/ 48	Generalization	Attributes Generalization	Facebook	(1): V, E, X(attributes), Yk(labels of known users) (2): Core, Ɛ (utility threshold)	(1): Yu (labels of unknown users) (2): Anonymized Graph
[62]	2017/ 63	Generalization	Weight Generalization	Facebook, CA-CondMat, Enron, Douban	(1): G(u), G(v), DF (different damping factors) (2): Graph Groups	(1): cost (Ge(u), Ge(v)) (2): Anonymized Graph
[58]	2018/ 5	Differential privacy	Hierarchical random graph (HRG)	Facebook, Enron and ca-HepPh	(1): original graph (2): subgraph, profile size, privacy parameter	(1): the approximate maximum independent set (2): HRG profile (Anonymized Graph)
[51]	2019/ 1	Differential privacy	Combining the dK-1, dK2, and dK-3 series together (noise on the dK-2)	Not mentioned	(1): dK-1 (2): dK-1, dK-2, dK-3	(1): the perturbed graph (2): Anonymized Graph
[52]	2019/ 1	Differential privacy	Injecting noise into the community and creating disturbances between them	The WebKB, the Citation and the Cora	(1): original graph (2): subgraph, m: profile size, p: privacy parameter	(1): the approximate maximum independent set (2): HRG profile
[53]	2019/ 0	Machine learning classification	Decision tree, Naïve Bayes, IBK, NB tree, and Bayes Network	Facebook	(1): Original graph	(1): the perturbed graph
[54]	2019/ 0	Differential privacy	Adding noise nodes to achieve k-degree	Facebook	(1): original graph G, Sequence Degree & Degree Groups	(1): Anonymized Graph
[55]	2019/ 0	Differential privacy (random alteration)	Random matrix (adding a small amount of noise)	Facebook, Live Journal and Pokec	(1): symmetric adjacency matrix, the number of random projections and variance for random noise	(1): Anonymized Graph
[56]	2019/ 3	Generalization	Clustering	Adult dataset	(1): G, Cluster Ci	(1): Anonymized Graph

[57]	2019/01	Generalization/ Differential privacy	All-Distance Sketch (ADS)	ca-HepTh, Facebook, and Enron	(1): Original graph G (2): Sketched graph Gs, change rate pc, |Ea|	(1): Sketched graph Gs, newly added edge |Ea| (2): Anonymized Graph

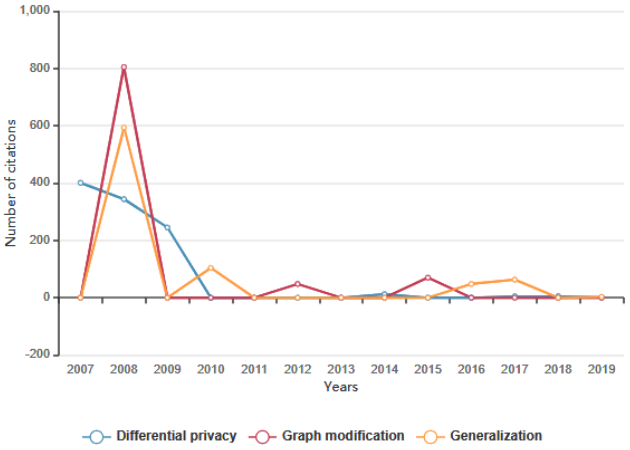


Fig. 2. Numbers of citations of the articles using Differential privacy, graph modification and generalization techniques depending on the Years.

The Line representation above represent the number citations according to the different researches about social networks anonymization using the graph modification, generalization and the differential privacy techniques from 2007 to the present. As we could estimates, the researches are still in development. So far, no method has guaranteed 100% the privacy and information loss. However, case studies and published research have shown how difficult it is to create a truly anonymous dataset while retaining enough underlying information for the needs of the task at hand, that why it is important to know or to have an idea about the main strengths and weaknesses of each technique can therefore be useful in deciding how to design an adequate anonymization process in a given context

REFERENCES

[1] J. Clement, "Number of daily active Facebook users worldwide as of 2nd quarter 2019," 14 Aug 2019. [Online]. Available: https://www.statista.com/statistics/346167/facebook-global-dau/.

[2] K. W. A. W.-C. F. a. P. S. Y. Benjamin C. M. Fung, Introduction to privacy-preserving data publishing: concepts and techniques, 2010.

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

[4] J. G. D. K. M. V. A. Machanavajjhala, "l-diversity: Privacy beyond kanonymity," in Data Engg, Atlanta, GA, USA, USA, 2006.

[5] S. G. Prakash M, "A new model for privacy preserving sensitive data mining," in 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12), Coimbatore, India, 2012.

[6] C. C. Mehmet Ercan Nergiz, "δ-presence without complete world knowledge," in IEEE Transactions on Knowledge and Data Engineering, 2009.

[7] Y. M. Sean J. Hickman, "Method and system for data pattern matching, masking and removal of sensitive data," 27 June 2013.

[8] T. M. M. T. a. W. T. Yang Xu, "A survey of privacy preserving data publishing using generalization and suppression," Applied Mathematics & Information Sciences 8, May 2014.

[9] J. M. Stephen E. Fienberg, "Data swapping: Variations on a theme by Dalenius and Reiss," in International Workshop on Privacy in Statistical Databases, 2005.

[10] R. Brand, "Microdata protection through noise addition," Inference Control in Statistical Databases, vol. 2316, pp. 97-116, 23 April 2002.

[11] G. S. Nelson, "Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification," 2015.

[12] P. D. P. C. Singapore, "Guide To Basic Data Anonymization Techniques," 2018.

[13] M. S. S. J. a. A. M. B.K. Tripathy, Privacy and Anonymization in Social Networks, vol. 65, 2014.

[14] L. S. Pierangela Samarati, "rotecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," P. Technical report, SRI International, 1998.

[15] G. M. D. J. a. D. T. M. Hay, "Resisting structural identification in anonymized social networks," in 34th International Conference on Very Large Databases (VLDB'08), 2008.

[16] G. M. D. J. P. W. S. S. Michael Hay, "Anonymizing social networks," Computer Science Department Faculty, p. 180, 2007.

[17] T. E. Liu K, "Towards identity anonymization on graphs," in ACM SIGMOD international conference on management of data, New York, NY, USA, 2008.

[18] W. X. Ying X, "Randomizing social networks: a spectrum preserving approach," in SIAM conference on data mining (SDM), Atlanta, Georgia, USA. SIAM, 2008.

[19] J. W. J. L. J. Z. Lian Liu, "Privacy preserving in social networks against sensitive edge disclosure," KY, 2008.

[20] E. Z. a. L. Getoor, "Preserving the privacy of sensitive relationships in graph data," in 1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD'07), 2007.

[21] T. M. T. Alina Campan, "A Clustering Approach for Data and Structural Anonymity," in In Proceedings of the 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD'08), in Conjunction with KDD'08, Las Vegas, Nevada, USA, 2008.

[22] J. P. Bin Zhou, "Preserving privacy in social networks against neighborhood attacks," in IEEE international conference on data engineering (ICDE), IEEE Computer Society, Washington, DC, USA, 2008.

[23] Z. a. J. Pei, "The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks," in Knowledge and information systems, 2010.

[24] D. S. T. Y. Q. Z. Graham Cormode, "Anonymizing bipartite graph data using safe grouping," in Proceedings of the 34th International Conference on Very Large Databases (VLDB'08), 2008.

[25] J. Kleinberg, "Challenges in Mining Social Network Data: Processes," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, USA, 2007.

[26] V. S. Arvind Narayanan, "De-anonymizing social networks," no. pp 173-187, 19 march 2009.

[27] F. L. X. Z. J. W. Wei Peng, "A Two-Stage Deanonymization Attack against Anonymized Social Networks," vol. 63, no. 2, pp. 290 - 303, february 2014.

[28] W. L. N. Z. G. P. M. a. R. B. Shouling Ji, "On your social network deanonymizablity: quantification and large scale evaluation with seed," in Proceedings of 2015 Symposium on Network and Distributed System Security (NDSS), 2015.

[29] a. V. K. V. Jyothi Vadisala, "Challenges in Social Network Data Privacy," International Journal of Computational Intelligence Research, vol. 13, pp. 965-979, 2017.

[30] J. P. W.-S. L. Bin Zhou, "A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data," ACM SIGKDD Explorations Newsletter, vol. 10, no. 2, pp. 12-22, December 2008.

[31] J. G. R. S. Alexandre Evfimievski, "Limiting privacy breaches in privacy preserving data mining," New York, NY, USA, 2003.

[32] R. S. D. T. Rakesh Agrawal, "Privacy preserving olap," in Proceedings of the 2005 ACM SIGMOD international conference on Management of data, New York, NY, USA, 2005.

[33] X. X. J. L. D. Z. Yufei Tao, "On anticorruption privacy preserving publication," in In Proceedings of the 24th International Conference on Data Engineering (ICDE'08), 2008.

[34] X. W. X Ying, "Randomizing Social Networks: a Spectrum Preserving Approach," in Proceedings of the 2008 SIAM International Conference on Data Mining, SIAM, Atlanta, 2008.

[35] A. G. T. T. Francesco Bonchi, "Identity obfuscation in graphs through the information theoretic lens," in Conf. on Data Engineering, IEEE, Washington, DC, 2011.

[36] S. Y. B. S. Lu X, "Fast identity anonymization on graphs," in 23rd international conference on database and expert systems applications (DEXA '12), Springer, Vienna, Austria, 2012.

[37] K. B. R. G. S. G. T. A. V. S. Chester S, "k-Anonymization of social networks by vertex addition," in ADBIS 2011 research communications, Vienna, Austria, 2011.

[38] G. J. S. U. V. S. Chester S, "Anonymizing subsets of social networks with degree constrained subgraphs," in IEEE international conference on advances on social networks analysis and mining (ASONAM), Washington, DC, USA, IEEE Computer Society, 2012.

[39] P. S. Y. X. K. Y. F. Chongjing Sun, "Privacy preserving social network publication against mutual friend attacks," in 2013 IEEE 13th International Conference on Data Mining Workshops, 2013.

[40] L. C. M. T. Ö. Lei Zou, "K-Automorphism: a general framework for privacy preserving network publication," Proceedings of the VLDB Endowment, vol. 2, no. 1, pp. 946-957, 2009.

[41] A. W.-c. F. J. L. James Cheng, "K-isomorphism: privacy preserving network publication against structural attacks," in International Conference on Management of data, Indianapolis, Indiana, USA, 2010.

[42] T. M. T. Alina Campan, "Data and Structural k-Anonymity in," in 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD'08), in Conjunction with KDD'08, Las Vegas, Nevada, USA, 2008.

[43] C. L. M. G. J. D. Michael Hay, "Accurate Estimation of the Degree Distribution of Private Networks," in Ninth IEEE International Conference on Data Mining, Miami, FL, USA, 2009.

[44] K. N. S. R. A. S. Shiva Prasad Kasiviswanathan, "Analyzing Graphs with Node Differential Privacy," in 10th theory of cryptography conference on Theory of Cryptography, Tokyo, Japan, 2013.

[45] S. R. A. S. Kobbi Nissim, "Smooth sensitivity and sampling in private data analysis," in Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, San Diego, California, USA, 2007.

[46] Y. S. S. B. Xue song Lu, "Fast Identity Anonymization on Graphs," in International Conference on Database and Expert Systems Applications, 2012.

[47] C. H. A. N. Sepp Hartung, "Improved Upper and Lower Bound Heuristics for Degree Anonymization in Social Networks," in International Symposium on Experimental Algorithms, 2014.

[48] G. M. D. J. D. T. P. W. Michael Hay, "Resisting structural re-identification in anonymized social networks," Proceedings of the VLDB Endowment, vol. 1, no. 1, pp. 102-114, 2008.

[49] P. K. W. X. G. L. Ying X, "Comparisons of randomization and K-degree anonymization schemes for privacy preserving social network publishing," in Proceedings of the 3rd workshop on social network mining and analysis, Paris, France, 2009.

[50] C.-R. J, "Privacy-preserving on graphs using randomization and edge-relevance," in Vicenç T (ed) International conference on modeling decisions for artifcial intelligence (MDAI). Springer International Publishing, Tokyo, 2014.

[51] F. L. Tianchong Gao, "Sharing Social Networks Using a Novel Differentially Private Graph Model," in 16th IEEE Annual Consumer Communications & Networking Conference (CCNC), 2019 .

[52] Y. X. Q. J. Y. T. Y. G. L.-e. W. X. L. Peng Liu, "Local differential privacy for social network publishing," Neurocomputing, 24 April 2019.

[53] M. S. T. S. Ratandeep Kaur, "Ensemble Based Model for Privacy in Online Social Network," in Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Jaipur - India, 2019.

[54] S. M. M. Seyedhashem Hamzehzadeh, "ANNM: A New Method for Adding Noise Nodes Which are Used Recently in Anonymization Methods in Social Networks," Wireless Personal Communications, 2019.

[55] A. X. L. R. J. Faraz Ahmed, "Publishing Social Network Graph Eigen-Spectrum with Privacy Guarantees," IEEE Transactions on Network Science and Engineering, 06 March 2019 .

[56] J. S. J. C.-R. Miguel Ros-Martín, "Scalable non-deterministic clustering-based k-anonymization for rich networks," International Journal of Information Security, vol. 18, no. 2, p. 219–238, April 2019.

[57] F. L. Tianchong Gao, "Privacy-Preserving Sketching for Online Social Network Data Publication," in 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), Boston, MA, USA, USA, 2019.

[58] F. L. Y. C. X. Z. Tianchong Gao, "Local Differential Privately Anonymizing Online Social Networks Under HRG-Based Model," IEEE Transactions on Computational Social Systems, vol. 5, pp. 1009 - 1020, 14 November 2018.

[59] G. M. D. J. D. T. C. L. Michael Hay, "Resisting structural re-identification in anonymized social networks," The International Journal on Very Large Data Bases, vol. 19, no. 6, pp. 797-823, 2010.

[60] Y. Z. J. C. J. S. M. T. Y. T. A. A.-D. M. A.-R. Tinghuai Ma, "KDVEM: a k-degree anonymity with vertex and edge modification algorithm," Computing, vol. 97, p. 1165–1184, 2015.

[61] Z. H. X. G. Y. L. Zhipeng Cai, "Collective Data-Sanitization for Preventing Sensitive Information Inference Attacks in Social Networks," IEEE Transactions on Dependable and Secure Computing, vol. 15, no. 4, pp. 577 - 590, 2016.

[62] G. W. F. L. S. Y. J. W. Qin Liu, "Preserving Privacy with Probabilistic Indistinguishability in Weighted Social Networks," IEEE Transactions on Parallel and Distributed Systems, vol. 28, p. 1417–1429, 20