2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, September 24-27, 2019

Improving Privacy in Graphs Through Node Addition

Nazanin Takbiri
Electrical and Computer Engineering
UMass-Amherst
ntakbiri@umass.edu

Xiaozhe Shao
Electrical and Computer Engineering
UMass-Amherst
xiaozheshao@engin.umass.edu

Lixin Gao
Electrical and Computer Engineering
UMass-Amherst
lgao@engin.umass.edu

Hossein Pishro-Nik
Electrical and Computer Engineering
UMass-Amherst
pishro@ecs.umass.edu

***Abstract*—The rapid growth of computer systems which generate graph data necessitates employing privacy-preserving mechanisms to protect users' identity. Since structure-based de-anonymization attacks can reveal users' identity's even when the graph is simply anonymized by employing naïve ID removal, recently, $k-$anonymity is proposed to secure users' privacy against the structure-based attack. Most of the work ensured graph privacy using fake edges, however, in some applications, edge addition or deletion might cause a significant change to the key property of the graph. Motivated by this fact, in this paper, we introduce a novel method which ensures privacy by adding fake nodes to the graph.**

First, we present a novel model which provides $k-$anonymity against one of the strongest attacks: seed-based attack. In this attack, the adversary knows the partial mapping between the main graph and the graph which is generated using the privacy-preserving mechanisms. We show that even if the adversary knows the mapping of all of the nodes except one, the last node can still have $k-$anonymity privacy.

Then, we turn our attention to the privacy of the graphs generated by inter-domain routing against degree attacks in which the degree sequence of the graph is known to the adversary. To ensure the privacy of networks against this attack, we propose a novel method which tries to add fake nodes in a way that the degree of all nodes have the same expected value.

***Index Terms*—Graph data, Autonomous System (AS)-level graph, Inter-domain routing, Privacy-Preserving Mechanism (PPM), anonymization and de-anonymization, structural attack, Seed-based attack, $k-$anonymity, $k-$automorohism, $k-$isomorphism.**

I. Introduction

Nowadays, a huge amount of data is generated from various different computer systems which can be modeled by a graph data. Social network data [1]–[4], communication data [5], Internet peer-to-peer networks and other network topologies [6], [7], mobility traced-based contact data [8] are some examples of computer systems and services which generate graph data. In these graphs, nodes represent users/systems and edges represent relationship between users/systems [9].

The necessity of sharing graph data for research purposes, data mining task, and commercial applications [10] presents a significant privacy threat to the users/systems [11] — even when the graph is simply anonymized — since the adversary can leverage their side-information about the structural graph to infer the private information of the users/systems which generated the graph [11]–[15].

The structure-based attacks have been introduced to graph data by [11], [12]. The structure-based attacks are aimed to de-anonymize anonymized users in terms of their uniquely distinguishable structural characteristics. There are different kinds of structure-based attacks that can be mainly categorized in the following groups:

- **Degree Attacks:** Assume the adversary knows the degree sequence of the graph, thus the adversary can use the degree sequence of the graph to uniquely identify one user/system if its degree is unique [16], [17]. $k-$degree anonymity is proposed by [18] in order to protect graph against this specific attack in a way that for each node, there exist at least $k-1$ other nodes with the same degree.
- **1-Neighborhood Attacks:** Assume the adversary knows the immediate neighbors of the target users, so they have complete information about the nodes adjacent to the target node. $k-$neighborhood anonymity is proposed by [19] to protect graph against the adversary who has knowledge about the neighborhood of the target node. In this privacy mechanism, for each node, there exist at least $k-1$ other users who have same neighborhood. [20], [21] also extend the previous work, to design an algorithm to defend against the adversary who has knowledge about the $d-$neighborhood of a target node.
- **Sub-graph Attacks:** The sub-graph attack in which the adversary knows a sub-graph around the target node is the general case of $1-$neighborhood attack. [12], [22], [23] proposed a method to protect users' identity from this specific kind of attack.
- **Hub-Fingerprint Attacks:** Assume there exist some hubs with high degree and high betweenness centrality [24] in the graph which have been identified in the released network. Now, assume there exists an adversary who has knowledge about distance between these sets of designated hub nodes and a target node, thus they can

This work was supported by National Science Foundation under grants CCF–1421957, CNS-1525836, CNS–1739462, CNS-1815412, and CCF–1918187.

use this knowledge to break the privacy of the target node.

Zou et al. [25] proposed the concept of $k-$automorphism in a way that can provide privacy against all of the above-mentioned structure-based attacks. Otherwise stated, a graph is $k-$automorphic, if for any node in the graph, there exist $k-1$ symmetric nodes based on any structural information such as their degree, their neighborhood, their distance from hubs, etc. The utility of this method characterized by using the number of faked edges added to the graph, thus, this method is specifically useful when the networks have symmetry properties [26]–[28]. However, the privacy mechanism proposed by [25] didn't address the privacy of the sensitive relationship between nodes, and in other words, the graph can suffer from path length leakage and edge leakage. [29], [30] showed the path length leakage and edge leakage exist even when a graph is $k-$anonymous and $k-$automorphic. Cheng et al. [31] proposed a new $k-$isomorphism privacy preserving mechanism which preserves the privacy of not only nodes but also edges. [31] convert the graph into k disjoint isomorphic sub-graphs and proved each node and each edges can be identified with the probability of $\frac{1}{k}$, thus they satisfy $k-$anonymity. However, the privacy mechanism proposed by [31] decreases the utility of the released graph since the edges between sub-graphs are deleted. In order to compensate this utility loss, Yang et al. [32] proposed a graph anonymization method in which the anonymous graph should satisfy $AK-$secure privacy preserving mechanism to minimize utility loss.

Today, knowledge of the adversary is not just limited to the structural of the graph, but also richer side-information in the form of seeds is available to them; otherwise stated, the adversary knows the mapping between the original graph and the anonymized graph for a subset nodes [17], [33]–[39]. Access of adversary to this side-information, which is difficult to control, make the previous anonymization technique more vulnerable. [17], [33]–[39] assume the network graph is generated using Erdös-Rényi random graph model [40], which is not a realistic assumption. [15], [41], proposed the first theoretical quantification of the perfect de-anonymization of a general setting in which the graph can be generated using any random model.

The bulk of previous work ensured privacy by deleting or adding fake edges. In this paper, we turn our attention to the case that privacy is guaranteed by adding fake nodes. Adding fake nodes could be a promising scheme to defend the graph privacy against the adversaries with side-information. Especially, it is necessary to add fake nodes in cases where other methods, such as node deletion, edge deletion and fake edge addition, might significantly decrease the utility of the released graph. For example, to study the inter-domain routing in the Internet, the Internet topology is usually modeled as an Autonomous System-level (AS-level) graph, where each node represents an Autonomous System (AS) which is a network operated by an institution and an edge between two nodes represents that two networks are directly connected [42]–[47]. In this scenario, the reliability property of a network to the rest of the Internet and the best path from one network to another are essential for the study of the inter-domain routing [7], [48]. Node or edge deletion might change the best path from one network to another or even make some networks unreachable from the rest of networks. Similarly, adding a fake edge between two real networks also changes the reliability property, since the fake edge leads to an additional path between two real networks.

In this paper, we proposed a novel method called $k-$fold replication method which preserves privacy of systems/users against seed-based attacks through adding fake nodes to the graph. Then, we turn our attention to the privacy of the graphs generated by inter-domain routing against degree attacks in which the degree sequence of the graph is known to the adversary. To address this issue, we propose a novel method which tries to add fake nodes in a way that the degree of all nodes have the same expected values.

The rest of the paper is organized as follows. In Section II, we present the general setting for privacy on graphs: system model, metrics, and definitions. Then, the conditions for achieving privacy by adding fake nodes in the case of seeded-based attack is discussed in Section III. In Section IV, we discuss how to ensure privacy for the networks using node addition, and in Section V, we conclude from the results.

II. A General Setting for Privacy on Graphs

In a general setting, we are given a graph $G = G(V, E)$, and without loss of generality, we write $V = \{1, 2, \cdots n\}$. Given a privacy mechanism $\mathcal{M}$ which is employed to guarantee privacy, a new graph $G^p(V^p, E^p)$ has been produced. In the simplest case, we might have $G^p \simeq G$ or in more advanced settings, we could construct G^p from G by adding fake vertices or adding/deleting edges.

The way G^p is constructed from G depends on the privacy mechanism $\mathcal{M}$ which is designed for the specific context, the constraints and requirements of the problem scenario. There exists a mapping function, $\sigma : V \mapsto V^p$ which is a function that determines the mapping between vertices of G and G^p. More specifically, for each vertex $v \in V$, $\sigma(v) \in V^p$ is the corresponding vertex in G^p. The adversary tries to identify users' identities by finding the mapping function σ.

Although we assume the privacy mechanism is known to adversary; the construction of G^p normally involves a randomized component and this randomness is what ensures the privacy. Here, we are interested in guaranteeing a privacy level, and our goal is employing a privacy preserving method to perturb the original graph structure to protect users' privacy while preserving as much data utility as possible.

We first briefly review the terminology that we use in this paper. Note that we adopt the definitions of $k-$automorphism and $k-$isomorphism from [25], [31], [49], [50], respectively.

Definition 1. *Graph Isomorphism* [49]: Given two graphs $G = (V^G, E^G)$ and $Q = (V^Q, E^Q)$, graph Q is isomorphic to

graph G if there exists a permutation function ($\Pi : V^Q \mapsto V^G$) such that (u, v) is in the set of graph edges E^G iff $(\Pi(u), \Pi(v))$ is in the set of graph edges E^Q.

Definition 2. *$k-$Isomorphism* [31]: A graph $G(V^G, E^G)$ is $k-$isomorphic if graph G consists of k disjoint subgraphs, i.e., $G = G_1 \cup G_2 \cup \cdots, \cup G_k$, where G_i and G_j are isomorphic for $i \neq j$.

Definition 3. *Graph Automorphism* [50] :Given graph $G = (V^G, E^G)$, it is a graph automorphism from graph G to itself if there exists an automorphic function ($\Pi : V^G \mapsto V^G$) such that (u, v) is in the set of graph edges E^G iff $(\Pi(u), \Pi(v))$ is in the set of graph edges E^G.

Definition 4. *$k-$Automporphism* [25]: A graph $G(V, E)$ is $k-$automorphic if for any node v in the graph, there exist $k-1$ different automorphic functions.

Here, we assume a strong adversary which employs seed-based attack which is defined as:

Definition 5. *Seed-Based Attack:* In this attack, we assume the adversary knows: (1) The main Graph (G) (or part of it), (2) The graph which is generated by privacy mechanism $\mathcal{M}$ (G^p), and (3) Partial of mapping function (σ), more specifically, the adversary knows the values of σ for a subset $V^s \subset V$. The goal of the adversary is to determine the values of σ for some vertices in $V - V^s$.

Now, before our method is discussed in detail, the measures of privacy cost and degree of anonymity that we employ are proposed.

Definition 6. *Privacy Cost:* The cost of a privacy mechanism is usually formulated as the distance measure between G and G^p. In other words, the more we distort G to make G^p, the more privacy cost we incur.

Privacy metrics could also depend on the situation. One type of privacy metric can be defined in terms of the minimum number of vertices that are needed to be revealed to the adversary ($|V^s|$), so that the adversary can recover the values of σ for some vertices in $V - V^s$. More specifically, we can have the following definition.

Definition 7. *Privacy Tolerance of Node v:* A privacy mechanism has a *privacy tolerance* τ_v for a vertex $v \in V$, if the adversary is unable to recover $\sigma(v)$ unless $|V^s| > \tau_v$. The largest value of τ_v that satisfies this property, is said to be the maximum tolerance of the mechanism for vertex v and we write

$$\mathcal{T}_v(\mathcal{M}) = \tau_v.$$

The value $\mathcal{T}_v(\mathcal{M})$ is specific to a vertex and depends on the structure of G, so we can provide the following measure for privacy of the mechanism that does not depend on the graph G.

Definition 8. *Privacy Tolerance:* A privacy mechanism has a *privacy tolerance* τ if for all graphs G with $|V| = n$, and all the vertices $v \in V^s$, we have $\mathcal{T}_v(\mathcal{M}) \geq \tau$. The largest value of τ that satisfies this property, is said to be the maximum privacy tolerance of the mechanism and we write

$$\mathcal{T}(\mathcal{M}) = \tau.$$

Clearly, we have $\mathcal{T}(\mathcal{M}) \leq n - 1$. Maximum tolerance gives some measure of privacy, but it does not provide all the needed information. More specifically, it does not give us a measure of the uncertainty of the adversary when $|V^s| < \mathcal{T}_n(\mathcal{M})$.

Now, we define *privacy function* that provides a much more complete picture about the privacy level of a mechanism. As we will see, the maximum tolerance defined above can be easily extracted from the privacy function. Intuitively, the privacy function, $h_n(\Lambda)$, gives us the guaranteed uncertainty about $\sigma(v)$, when the adversary has labels of Λ vertices (i.e., $|V^s| = \Lambda$).

We now provide the formal definition of privacy function for a privacy mechanism $\mathcal{M}$. Entropy and mutual information usually provide an effective tools for defining privacy measures [51], [52]. If $|V^s| = \Lambda$, we write

$$V^s = \{u_1, u_2, \cdots, u_\Lambda\}.$$

Definition 9. *Privacy Function:* For a privacy mechanism $\mathcal{M}$, privacy function which is denoted as $h_n : \{0, 1, 2, \cdots, n\} \mapsto \mathbb{R}^+$ is defined as follows:

$$h_n(\Lambda) = \min \Big\{ H\big(\sigma(v) \big| \sigma(u_1), \sigma(u_2), \cdots, \sigma(u_\Lambda)\big) : |V| = n, |V^s| = \Lambda, v \in V - V^s \Big\},$$

where $H(\cdot|\cdot)$ denotes the conditional entropy. Note that [53]–[57] also employed entropy to define degree of anonymity achieved by the users of a system towards particular attackers.

Note that the minimum is taken over all graphs G with n vertices, all $v \in V$, and all $V^s \subset V$. With this definition, it easy to see

$$\mathcal{T}_n(\mathcal{M}) = \min\{\tau : h_n(\tau) = 0\}.$$

III. Achieving Privacy Against Seed-Based Attacks

When the Internet is modeled as an AS-level graph, to preserve the utility of the AS-level graph, the anonymization scheme should maintain the key properties, such as the reachability and reliability between networks. In the studies of the inter-domain routing, it could be essential to figure out the reliability and the best path from one network to another [7], [48]. For example, to derive the global routing table, the best route of each network should be derived based on the AS-level graph [7].

In an AS-level graph, adding fake nodes into the AS-level graph introduces additional fake networks. After that, fake edges can be added between two fake nodes or between a fake node and a real node. These fake edges represent fake

connections between two networks. Note, we do not add fake edges between two origin nodes, since it decreases the utility of the graph.

Adding fake nodes can preserve the utility of an AS-level graph, since the key properties of the graph can be maintained. On one hand, the additional fake nodes do not remove the original paths between two networks in the graph. Therefore, if a path from a real network to another real network exists in the original graph, the path always exists in the generated graph. On the other hand, even if the additional fake nodes and edges might lead to additional paths from a network to another, the the export policy of fake nodes can not altered to guarantee that these additional paths between two real networks are invalid paths in terms of inter-domain routing. That is, a fake network will not announce a route that goes through real networks to its neighboring nodes that represent real networks. Therefore, a path with any fake edge will not be a valid route from one real network to another real network.

In this section, we propose approaches to make a graph $k-$anonymized through adding fake nodes. Here, the adversary has side-information not only about the structural information of the graph, but also in the form of seeds. Now, since we focus on privacy mechanisms that add fake nodes to preserve privacy, we use the number of fake nodes to measure the cost of a privacy mechanism as follows.

Definition 10. *Privacy Cost:* Given an original graph G and the generated graph by privacy mechanism G^p, the privacy cost is defined as the number of fake nodes added to the original graph, in other words,

$$Cost\left(G, G^p\right) = |V^p| - |V|$$

.

A. The Naïve Approach

In this method, we generate exactly $k-1$ copies of the original graph to achieve a new graph which satisfies $k-$automorphism/$k-$isomorphism.

As it is shown in Figure 1, the set of nodes of the generated graph can be defined as:

$$V^p = \left\{(i,j) : i \in \{1,2,\cdots,k\}, j \in \{1,2,\cdots,n\}\right\},$$

thus, the privacy cost can be calculated as:

$$Cost(G, G^p) = (k-1)n.$$

Note that the privacy tolerance of this graph which is generated using Naïve approach is one, in other words, if only one of the nodes is known to the adversary, the adversary can recover the whole graph and break privacy of users.

$$\mathcal{T}_n(\mathcal{M}) = 1.$$

Now, by using the fact that the privacy tolerance of this generated graph is equal to one, we can conclude the privacy function of the Naïve approach can be calculated as:

$$h_n(\Lambda) = \begin{cases} \log_2 k, & \text{for } \Lambda = 0. \\ 0, & \text{for } \Lambda \geq 1. \end{cases}$$

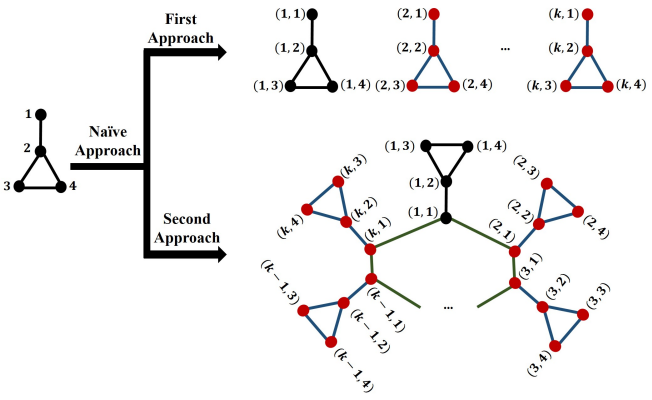

Fig. 1: An Example which demonstrates how Naïve approach is employed to satisfy $k-$anonymity. First approach satisfies $k-$isomorphism by generation $k-1$ copies of the original graph. Second approach satisfies $k-$automorphism by connecting all of the $k-1$ copied graphs.

As a result, the "$k-$anonymity" type approaches to privacy are not usually sufficient.

B. The Replication Method

Motivated by the above discussions, here we introduce a technique to guarantee privacy against seed-based attacks. We call it the replication method, as it can be thought of as replicating the vertices of the original graph in a special way. The basic idea is as follows. Start from any vertex in the graph $v_1 \in G$ and add a new vertex to the graph which is connected to all the neighbors of v_1. Call the new graph G_1. Now identify another vertex in $v_2 \in G$ and add a new vertex in G_1 that is connected to all neighbors of v_2 in G_1. This will give you G_2. Repeat this process until you exhaust all vertices of G. At the end you will obtain G_n which will be our graph G^p. This is a two-fold replication method. You can simply extend this to a k-fold replication by repeating the whole process $k-1$ times. Below we formally introduce the technique and show that it provides a high guarantee against seed-based attacks.

The set of nodes of the graph which is generated by $k-$fold replication can be defined as:

$$V^p = \Big\{(i,j) : i \in \{1,2,\cdots,k\}, j \in \{1,2,\cdots,n\}, \\ \big((i,j),(u,v)\big) \in V^p \text{ iff } (j,v) \in V\Big\},$$

thus, the privacy cost can be calculated as:

$$Cost(G, G^p) = (k-1)n,$$

which is the same as the Naïve method. The number of fake edges needed in the $k-$replicated method can be also calculated as:

$$\Delta|E| = |E^p| - |E| = (k^2-1)|E|.$$

Note that the privacy tolerance of the graph $\mathcal{T}_n(\mathcal{M})$ can be calculated as:

$$\mathcal{T}_n(\mathcal{M}) = n-1,$$

in other words, if the adversary knows the mapping function (σ) for $n-1$ nodes–which is the maximum possible value for the privacy tolerance– they still identify the last node with the probability of $\frac{1}{k}$, so this method can obtain the maximum possible value for the privacy tolerance.

Theorem 1. For $k-$fold replication method, the privacy function is bigger than or equal to $\log_2 k$ for the case $\Lambda \in \{0, 1, \cdots, n-1\}$. In other words,

$$h_n(\Lambda) = \log_2 k \text{ for all } \Lambda = 0, 1, 2, \cdots, n-1.$$

Proof. In the first step, we prove that $h_n(\Lambda) \geq \log_2 k$. To do so, let's assume the adversary knows the mapping function for $n-1$ nodes, and there is only one unknown node. Since the replication method creates a k vertices with the same neighbors, by symmetry, the privacy function can be calculated as

$$\begin{aligned} h_n(n-1) &= -\sum_{i=1}^{k} p_i \log_2 p_i \\ &= -\sum_{i=1}^{k} \frac{1}{k} log_2 \frac{1}{k} \\ &= \log_2 k. \end{aligned} \tag{1}$$

Now, given the fact that conditioning reduces entropy, for all $\Lambda \leq n-1$, we have

$$H\left(\sigma(v)|\sigma(u_1), \sigma(u_2), \cdots, \sigma(u_\Lambda)\right) \geq H\left(\sigma(v)|\sigma(u_1), \sigma(u_2), \cdots, \sigma(u_n)\right),$$

and as a result, for all $\Lambda \leq n-1$,

$$\begin{aligned} h_n(\Lambda) &\geq h_n(n-1) \\ &= \log_2 k. \end{aligned} \tag{2}$$

Now, from (1) and (2), we can conclude the privacy function for the $k-$replication method satisfies

$$h_n(\Lambda) \geq \log_2 k \text{ for all } \Lambda = 0, 1, 2, \cdots, n-1.$$

In the second step, to show that

$$h_n(\Lambda) \leq \log_2 k \text{ for all } \Lambda = 0, 1, 2, \cdots, n-1,$$

it suffices to provide examples of scenarios (for all n and $\Lambda = 0, 1, 2, \cdots, n-1$) where

$$H\left(\sigma(v)\middle|\sigma(u_1), \sigma(u_2), \cdots, \sigma(u_\Lambda)\right) \leq \log_2 k.$$

Consider the graph $G(V, E)$, which is shown in Figure 2. In this graph, there exist $n-1$ node with degree of one, and one node with degree of $n-1$. Thus, the degree sequence of the original graph is equal to

$$\mathbf{d}^G = [\underbrace{1, 1, \cdots, 1}_{\text{n-1}}, n-1],$$

Note that the $k-$fold replication method, increases degree of each node by a factor of k, thus, the degree sequence of graph after $k-$fold replication is equal to:

$$\mathbf{d}^{G^P} = [\underbrace{k, k, \cdots, k}_{\text{k(n-1)}}, \underbrace{k(n-1), k(n-1), \cdots, k(n-1)}_{\text{k}}].$$

This means that in the $k-$fold replication method, the uncertainty of the adversary about the node which has degree of $n-1$ is always less than or equal to $\log_2 k$. Therefore,

$$h_n(\Lambda) \leq \log_2 k \text{ for all } \Lambda = 0, 1, 2, \cdots, n-1.$$

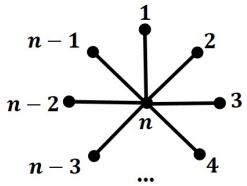

Fig. 2: A graph with $n-1$ node with degree of one, and one node with degree of $n-1$.

□

The idea behind $k-$fold replication method is explained by the following example:

Example 1. Figure 3 shows the replication method for the case $k = 2$; namely, two-fold replication. In this example, number of nodes (n) is equal to 3, thus the privacy cost is calculated as

$$Cost(G, G^p) = 3,$$

and the privacy tolerance of the graph is equal to

$$\mathcal{T}_n(\mathcal{M}) = 2.$$

Thus, the privacy function of this two-replicated method is equal $h_n(\Lambda) = 1$ for all $\Lambda = 0, 1, 2, \cdots, n-1$.

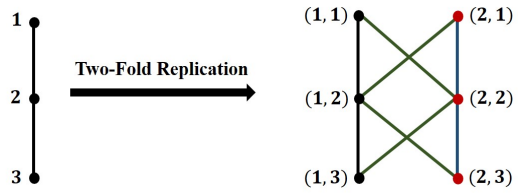

Fig. 3: Two-fold Replication Method. Black nodes represent real nodes and red nodes represent added fake nodes. Black edges represent the real edges, green edges represent the fake edges which connect a fake node and a real node, and blue edges represent the fake edges which connect two fake nodes.

IV. Privacy of Networks Through Node Addition

In this section, we want to provide privacy regarding the *Degree Attack*. Degree attack is one type of the structure-based attacks, in which, the adversary knows degree sequence of the graph, in other words, they know the degree of all nodes in the graph [18], [22].

In the scenario of the AS-level graph, the degree of each network can be inferred from publicly available datasets, such as Border Gateway Protocol (BGP) routing information provided by Route Views project [59] and Routing Information Service (RIS) provided by Réseaux IP Européens Network Coordination Center (RIPE NCC) [60]. The large networks at the core of the Internet, such as Tier-1 Internet Service Providers (ISPs), have high degree. Through their special degrees, the large networks can be easily mapped to the nodes in the AS-level graph. Then these identified networks can benefit the follow-up de-anonymization greatly, since these identified networks will provide a lot of structural information. When these networks are identified, the privacy information (routing policies) attached to the nodes will be disclosed.

Although the privacy schemes in Section III can provide privacy regarding various structure-based attacks, it incurs high privacy cost. Namely, $k-1$ replications for each real node are added to the original graph. In this section, we propose a more efficient privacy mechanism in terms of privacy cost against degree attack.

A. A General Setting for Privacy Against Degree Attack

The original graph such as power law graph is asymmetric in terms of degree distribution [61], which reveals a lot of information to the adversary who has knowledge about degree distribution of the graph data. Assume $G(V,E)$ denotes a graph data with set of nodes V ($|V|=n$) , and set of edges E. Without loss of generality, we assume $v \in V = \{1,2,\cdots,n\}$. Our main goal is protecting identity of all of the nodes of this graph from a strong adversary who has full knowledge about the degree sequence of this graph. Now, assume the adversary knows the $1\times n$ vector containing the degree of node v,

$$\mathbf{d} = [d_1, d_2, \cdots, d_n],$$

where d_v denotes degree of node v. In order to preserve the privacy of nodes, a new random graph called G^P is generate in a way that all vertices have the same expected degree. In order to have the same expected values for all the vertices, we add m fake nodes to the original graph (G). $U = \{1,2,\cdots,m\}$ denotes the set of fake nodes; thus, $u \in \{1,2,\cdots,m\}$. Now, new graph $G^p = (V^p, E^p)$ is constructed using probabilistic method to introduce more uncertainty to the model and as a result, confuse the adversary more. In this method, fake edges which connects two fake nodes or one fake node with one real node randomly; to be more specific,

- Each edge between a real node and a fake node is included in the graph with probability p_{vu} independent from every other edge, where $v \in \{1,2,\cdots n\}$ and $u \in \{1,2,\cdots,m\}$. There exist $n\times m$ different values for p_{vu}'s.
- Each edge between two fake nodes is included in the graph with probability q_{uw} independent from every other edge, where $u,w \in \{1,2,\cdots,m\}$. There exist $\binom{m}{2}$ different values for q_{uw}'s.

Now, as shown in Figure 4, for all real nodes $v \in \{1,2,\cdots,n\}$, the expected value of degree of each real node after this operation can be calculated as:

$$\mathbb{E}[D_v] = d_v + \sum_{u=1}^{m} p_{vu}. \tag{3}$$

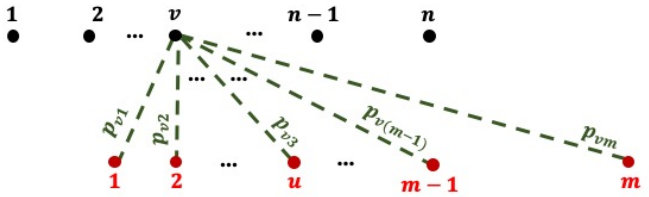

Fig. 4: The real node v and its degree after the anonymization operation is employed to generate the random graph G^P. In this figure, for simplicity, the real edges of the graph is not shown.

Also, as shown in Figure 5, for all fake nodes $u \in \{1,2,\cdots,m\}$, the expected value of degree of each fake node after this operation can be calculated as:

$$\mathbb{E}[D'_u] = \sum_{v=1}^{n} p_{vu} + \sum_{\substack{w=1 \\ w\neq u}}^{m} q_{uw}. \tag{4}$$

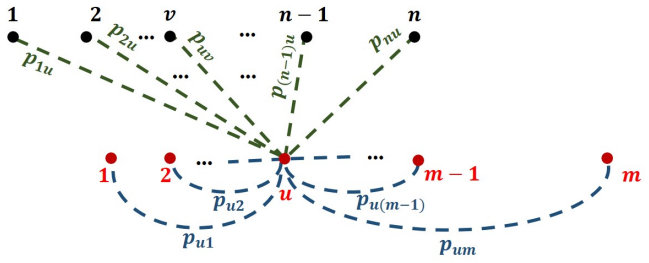

Fig. 5: The fake node u and its degree after the anonymization operation is employed to generate the random graph G^P.

Our goal is adjusting the value of p_{vu}'s and q_{uw}'s in a way that identity of users against degree attack will be protected.

Theorem 2. Consider a general graph $G(V,E)$. If all the followings hold

1) $m = n - \frac{2|E|}{a}$, where a is a constant number which is greater than the maximum degree of original graph.
2) $p_{vu} = \frac{a-d_v}{m}$, for all $v \in \{1,2,\cdots,n\}$ and $u \in \{1,2,\cdots,m\}$.
3) $q_{vw} = 0$, for all $u \in \{1,2,\cdots,n\}$ and $w \in \{1,2,\cdots,m\}$.

then, the expected values of all the real and fake nodes have the same value which is equal to a.

Proof. After adding m fake nodes which each of them is connected to real nodes with probability of $p_{vu} = \frac{a-d_v}{m}$, for all $v \in \{1, 2, \cdots, n\}$, the expected value of of degree of each real node can be calculated by using (3):

$$\begin{aligned}\mathbb{E}[D_v] &= d_v + m\left(\frac{a-d_v}{m}\right)\\ &= a,\end{aligned} \tag{5}$$

Also, for all $u \in \{1, 2, \cdots, m\}$, by using (4), we have

$$\begin{aligned}\mathbb{E}[D'_u] &= \sum_{v=1}^{n} p_{vu} + \sum_{\substack{w=1\\ w\neq u}}^{m} q_{uw}\\ &= \sum_{v=1}^{n} \frac{a-d_v}{m}\\ &= \frac{na-2|E|}{\frac{na-2|E|}{a}}\\ &= a,\end{aligned} \tag{6}$$

since each fake node is connected to a real node with probability of $p_{vu} = \frac{a-d_v}{m}$, and two fake nodes are connected with probability of $q_{uw} = 0$.

Now, from (5) and (6), we can conclude after this operation, the expected values of all the real and fake nodes have the same value which is equal to a. Now, since all the nodes have the same expected values, the adversary gets more confused. □

V. Conclusion

The wide presence of graph data which are generated by computer systems requires graph-based privacy-preserving mechanisms. Most of the proposed privacy-preserving mechanisms ensure privacy by deleting/ adding edges. However, adding or deleting edges between two real nodes might significantly decrease the utility of the released graph data, thus, in this paper, we presented graph-based privacy preserving mechanisms which ensure privacy by adding fake nodes to the original graph. In the first part of the paper, we proposed a novel mechanism called $k-$replication method to protect the identity of users against one of the strongest attacks called seed-based attack. In the second part of the paper, we improve privacy of inter-domain routing against degree attack by adding fake nodes and edges in a way that the degree of all nodes have the same expected values.

References

[1] L. Backstrom, D. P. Huttenlocher, J. M. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *KDD*, 2006.

[2] G. Kossinets, J. M. Kleinberg, and D. J. Watts, "The structure of information pathways in a social communication network," *ArXiv*, vol. abs/0806.3201, 2008.

[3] G. Bergami, F. Bertini, and D. Montesi, "On approximate nesting of multiple social network graphs: a preliminary study," in *23rd International Database Engineering and Applications Symposium (IDEAS19)*.

[4] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Link mining: models, algorithms, and applications*. Springer, 2010, pp. 337–357.

[5] L. V. S. Lakshmanan, R. T. Ng, and G. Ramesh, "To do or not to do: The dilemma of disclosing anonymized data," in *SIGMOD Conference*, 2005.

[6] M. Ripeanu and I. T. Foster, "Mapping the gnutella network: Macroscopic properties of large-scale peer-to-peer systems," in *IPTPS*, 2002.

[7] G. Asharov, D. Demmler, M. Schapira, T. Schneider, G. Segev, S. Shenker, and M. Zohner, "Privacy-preserving interdomain routing at internet scale," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 3, pp. 147–167, 2017.

[8] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: using social network as a side-channel," in *ACM Conference on Computer and Communications Security*, 2012.

[9] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[10] S. Ji, P. Mittal, and R. A. Beyah, "Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey," *IEEE Communications Surveys and Tutorials*, vol. 19, pp. 1305–1326, 2017.

[11] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *2009 30th IEEE Symposium on Security and Privacy*. IEEE, 2009, pp. 173–187.

[12] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 181–190.

[13] S. Ji, W. Li, M. Srivatsa, and R. A. Beyah, "Structural data de-anonymization: Quantification, practice, and implications," in *ACM Conference on Computer and Communications Security*, 2014.

[14] S. Ji, W. Li, M. Srivatsa, J. He, and R. A. Beyah, "Structure based data de-anonymization of social networks and mobility traces," in *ISC*, 2014.

[15] S. Ji, W. Li, N. Z. Gong, P. Mittal, and R. A. Beyah, "On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge," in *NDSS*, 2015.

[16] P. Pedarsani and M. Grossglauser, "On the privacy of anonymized networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1235–1243.

[17] L. Yartseva and M. Grossglauser, "On the performance of percolation graph matching," in *Proceedings of the first ACM conference on Online social networks*. ACM, 2013, pp. 119–130.

[18] K. Liu and E. Terzi, "Towards identity anonymization on graphs," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 93–106.

[19] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks." in *ICDE*, vol. 8. Citeseer, 2008, pp. 506–515.

[20] H. Jin, Z.-x. ZHANG, S.-c. LIU, and S.-g. JU, "Preserving privacy in social networks based on d-neighborhood subgraph anonymity," *Application Research of Computers*, no. 11, p. 88, 2011.

[21] H. Wu, J. Zhang, B. Wang, J. Yang, and B. Sun, "d, k-anonymity for social networks publication against neighborhood attacks," *Journal of Convergence Information Technology JCIT*, vol. 8, no. 2, pp. 59–67, 2013.

[22] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting structural re-identification in anonymized social networks," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 102–114, 2008.

[23] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, "Anonymizing social networks," *Computer science department faculty publication series*, p. 180, 2007.

[24] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[25] L. Zou, L. Chen, and M. T. Özsu, "K-automorphism: A general framework for privacy preserving network publication," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 946–957, 2009.

[26] J. Lauri and R. Scapellato, *Topics in graph automorphisms and reconstruction*. Cambridge University Press, 2016, vol. 432.

[27] H. Wang, G. Yan, and Y. Xiao, "Symmetry in world trade network," *Journal of Systems Science and Complexity*, vol. 22, no. 2, pp. 280–290, 2009.

[28] X. Ying and X. Wu, "Randomizing social networks: a spectrum preserving approach," in *proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 2008, pp. 739–750.

[29] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, 2006, pp. 24–24.

[30] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 754–759.

[31] J. Cheng, A. W.-c. Fu, and J. Liu, "K-isomorphism: privacy preserving network publication against structural attacks," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010, pp. 459–470.

[32] J. Yang, B. Wang, X. Yang, H. Zhang, and G. Xiang, "A secure k-automorphism privacy preserving approach with high data utility in social networks," *Security and Communication Networks*, vol. 7, no. 9, pp. 1399–1411, 2014.

[33] E. Onaran, S. Garg, and E. Erkip, "Optimal de-anonymization in random graphs with community structure," in *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 709–713.

[34] E. Kazemi, S. H. Hassani, and M. Grossglauser, "Growing a graph matching from a handful of seeds," *Proceedings of the VLDB Endowment*, vol. 8, no. 10, pp. 1010–1021, 2015.

[35] V. Lyzinski, D. E. Fishkind, and C. E. Priebe, "Seeded graph matching for correlated erdös-rényi graphs." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3513–3540, 2014.

[36] E. Kazemi, L. Yartseva, and M. Grossglauser, "When can two unlabeled networks be aligned under partial overlap?" in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2015, pp. 33–42.

[37] D. Cullina and N. Kiyavash, "Improved achievability and converse bounds for erdos-rényi graph matching," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 1. ACM, 2016, pp. 63–72.

[38] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," *ArXiv*, vol. abs/1307.1690, 2014.

[39] F. Shirani, S. Garg, and E. Erkip, "Seeded graph matching: Efficient algorithms and theoretical guarantees," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2017, pp. 253–257.

[40] E. Kazemi, "Network alignment: Theory, algorithms, and applications," EPFL, Tech. Rep., 2016.

[41] S. Ji, W. Li, N. Z. Gong, P. Mittal, and R. A. Beyah, "Seed-based de-anonymizability quantification of social networks," *IEEE Transactions on Information Forensics and Security*, vol. 11, pp. 1398–1411, 2016.

[42] L. Gao, "On inferring autonomous system relationships in the internet," *IEEE/ACM Transactions on Networking*, vol. 9, no. 6, pp. 733–745, Dec 2001.

[43] T. G. Griffin and G. Wilfong, "An analysis of bgp convergence properties," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, ser. SIGCOMM '99. New York, NY, USA: ACM, 1999, pp. 277–288. [Online]. Available: http://doi.acm.org/10.1145/316188.316231

[44] J. a. L. Sobrinho, "Network routing with path vector protocols: Theory and applications," in *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '03. New York, NY, USA: ACM, 2003, pp. 49–60. [Online]. Available: http://doi.acm.org/10.1145/863955.863963

[45] A. Sosnovich, O. Grumberg, and G. Nakibly, "Analyzing internet routing security using model checking," in *Proceedings of the 20th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning - Volume 9450*, ser. LPAR-20 2015. New York, NY, USA: Springer-Verlag New York, Inc., 2015, pp. 112–129. [Online]. Available: https://doi.org/10.1007/978-3-662-48899-7_9

[46] V. Giotsas, M. Luckie, B. Huffaker, and k. claffy, "Inferring Complex AS Relationships," in *Internet Measurement Conference (IMC)*, Nov 2014, pp. 23–30.

[47] Y. Wang, M. Schapira, and J. Rexford, "Neighbor-specific bgp: More flexible routing policies while improving global stability," in *Proceedings of the Eleventh International Joint Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '09. New York, NY, USA: ACM, 2009, pp. 217–228. [Online]. Available: http://doi.acm.org/10.1145/1555349.1555375

[48] R. Klöti, V. Kotronis, B. Ager, and X. Dimitropoulos, "Policy-compliant path diversity and bisection bandwidth," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, April 2015, pp. 675–683.

[49] M. De Santo, P. Foggia, C. Sansone, and M. Vento, "A large database of graphs and its use for benchmarking graph isomorphism algorithms," *Pattern Recognition Letters*, vol. 24, no. 8, pp. 1067–1079, 2003.

[50] L. W. Beineke, R. J. Wilson, P. J. Cameron *et al.*, *Topics in algebraic graph theory*. Cambridge University Press, 2004, vol. 102.

[51] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching anonymized and obfuscated time series to users' profiles," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 724–741, 2019.

[52] Z. Montazeri, A. Houmansadr, and H. Pishro-Nik, "Achieving Perfect Location Privacy in Wireless Devices Using Anonymization," *IEEE Transaction on Information Forensics and Security*, vol. 12, no. 11, pp. 2683–2698, 2017.

[53] C. Diaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *International Workshop on Privacy Enhancing Technologies*. Springer, 2002, pp. 54–68.

[54] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *International Workshop on Privacy Enhancing Technologies*. Springer, 2002, pp. 41–53.

[55] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, "Community-enhanced de-anonymization of online social networks," in *Proceedings of the 2014 acm sigsac conference on computer and communications security*. ACM, 2014, pp. 537–548.

[56] S. Ji, T. Du, Z. Hong, T. Wang, and R. Beyah, "Quantifying graph anonymity, utility, and de-anonymity," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1736–1744.

[57] S. Ji, "Evaluating the security of anonymized big graph/structural data," Ph.D. dissertation, Georgia Institute of Technology, 2016.

[58] J. L. Gross and J. Yellen, *Graph theory and its applications*. CRC press, 2005.

[59] U. of Oregon, "Route views." [Online]. Available: www.routeviews.org

[60] R. NCC, "Routing information service." [Online]. Available: https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris

[61] M. Hay, C. Li, G. Miklau, and D. Jensen, "Accurate estimation of the degree distribution of private networks," in *2009 Ninth IEEE International Conference on Data Mining*. IEEE, 2009, pp. 169–178.