De-SAG: On the De-anonymization of Structure-Attribute Graph Data

Shouling Ji, *Member, IEEE,* Ting Wang, Jianhai Chen, *Member, IEEE,* Weiqing Li, *Student Member, IEEE,* Prateek Mittal, *Member, IEEE,* and Raheem Beyah, *Senior Member, IEEE*

Abstract—In this paper, we study the impacts of non-Personal Identifiable Information (non-PII) on the privacy of graph data with attribute information (e.g., social networks data with users' profiles (attributes)), namely *Structure-Attribute Graph* (SAG) data, both theoretically and empirically. Our main contributions are two-fold: (i) we conduct the first *attribute-based anonymity analysis* for SAG data under both preliminary and general models. By careful quantification, we obtain the explicit correlation between the graph anonymity and the attribute information. We also validate our analysis through numerical and real world data-based evaluations and the results indicate that the non-PII can also lead to significant anonymity loss; and (ii) according to our theoretical analysis, we propose a new de-anonymization framework for SAG data, namely De-SAG, which takes into account both the graph structure and the attribute information to the best of our knowledge. By extensive experiments, we demonstrate that De-SAG can significantly improve the performance of state-of-the-art graph de-anonymization attacks. Our attribute-based anonymity analysis and de-anonymization framework are expected to provide data owners and researchers a more complete understanding on the privacy vulnerability of graph data, and thus shed light on future graph anonymization and de-anonymization research.

Index Terms—Anonymity analysis; de-anonymization; Structure-Attribute Graph (SAG) data; evaluation

1 INTRODUCTION

Different from traditional tabular/relational data, many data generated by modern computer systems are graph data consisting of nodes and links [1][2][3][4][5]. Normally, the nodes in the graphs represent users (or devices operated by users) and the links represent the relationships among users. In practice, typical graph data includes social networks data [1][2], communication data [4][21], network topology data [4][21], mobility traces [3][7], etc. For example, when modeling social networks data (e.g., Facebook [37], Twitter [38], and Google Plus (GP) [39]) as graphs, the nodes represent users and the links represent social relationships among users (e.g., friendships, follow relationships, and circle relationships).

As with traditional tabular/relational data, graph data are also very useful for many applications, e.g., academic research, data mining tasks, advertisements, commercial decision support, fraud/terrorist detection, the study of disease diffusion, and thus they are frequently shared/published to researchers, commercial parters, and/or even the public [2]-[7], [19]-[21]. On the other hand, graph data also carry a lot of sensitive private information of users who generate them (e.g., birthdate, salary, sexual orientation, political preference), which might be learned by adversaries. Therefore, graph data are usually anonymized before being shared/published [22][29][31].

However, as shown both theoretically and empirically [2], [4], [17], the anonymized graph data may still be susceptible to *structure-based de-anonymization attacks*, which de-anonymize an anonymized graph leveraging the structural (topological) similarity between the anonymized graph and an auxiliary graph. In [1], [2], [3], [6], [7], [15], [16], several *two-phase seed-based de-anonymization attacks* are proposed. In the first phase of the attack, seed mappings are identified from the anonymized graph to the auxiliary graph. Then, in the second phase, the de-anonymization is propagated to other anonymized users from the seed users leveraging various de-anonymization techniques. In [4], a *seed-free* (or blind) de-anonymization scheme is proposed, under which the de-anonymization is conducted via minimizing an error function defined on the structural difference between the anonymized graph and the auxiliary graph. In addition to the graph structure-based de-anonymization attacks, recently, the study of understanding and quantifying the *structure-based de-anonymizability* of graph data has also drawn a lot of attention [4][14][15][16][17].

In most real scenarios of sharing/publishing graph data, in addition to sharing/publishing the graph structure, a lot of non-*Personal Identifiable Information* (non-PII), or *attribute information*, associated with graph users is also shared or published, e.g., gender, education, city, country, interests [19]-[21]. Therefore, when studying anonymization and de-anonymization techniques for graph data, the following question can be posed: *what are the impacts of the attribute information on the anonymity/de-anonymizability of graph data?*

- *S. Ji is the corresponding author.*

- *S. Ji and J. Chen are with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China.*
  E-mail: sji@zju.edu.cn
- *T. Wang is with the Department of Computer Science, Lehigh University, Bethlehem, PA 18015, USA.*
  E-mail: ting@cse.lehigh.edu
- *W. Li and R. Beyah are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.*
  E-mail: wli64@gatech.edu, rbeyah@ece.gatech.edu
- *P. Mittal is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08540, USA.*
  E-mail: pmittal@princeton.edu

However, in existing graph data de-anonymization research [1]-[7][15][16], only graph structure information is considered. Similarly, existing graph anonymity quantification research [4][14][15][16][17] only considers the graph structure, which gives an incomplete picture of the actual privacy vulnerability of graph data. To address the aforementioned open problem, we study the impact of attribute information (non-PII) on the privacy of graph data both theoretically and empirically. In this paper, to distinguish between graph data with just graph structure and graph data with structure and attributes, we name the graph data with structure and attribute information *Structure-Attribute Graph* (SAG) data. Our main contributions can be summarized as follows.

1) We conduct the first *attribute-based anonymity analysis* of SAG data under both preliminary and general data models. By careful quantification, we explicitly demonstrate the correlation between the achievable graph anonymity and the attribute information. Our theoretical results demonstrate that the attribute information, even as non-PII, can also lead to significant anonymity loss of graph data. We also validate our analysis by both numerical evaluation and real world SAG data-based evaluation. The evaluation results further confirm our anonymity analysis. Our attribute-based anonymity analysis together with existing structure-based de-anonymizability quantifications provide data owners and researchers a more complete understanding of the privacy of graph data.
2) According to our attribute-based anonymity analysis, we propose a new de-anonymization attack on graph data, namely De-SAG, which takes into account both graph structure and attribute information to the best of our knowledge. Through extensive evaluations leveraging real world SAG data, we demonstrate that De-SAG can significantly enhance existing structure-based de-anonymization attacks. For instance, when de-anonymizing a Facebook dataset (4,039 users, 88,234 user-user links, 1,283 attributes, 37,257 user-attribute links), De-SAG has a 3.82 $\sim$ 10.1 times better de-anonymization performance than state-of-the-art structure-based de-anonymization attacks [4][16].

Roadmap. In the rest of this paper, we discuss related work in Section 2. In Section 3, we provide the data model, preliminaries, and definitions. The attribute-based anonymity analysis and evaluation are conducted in Section 4. Then, we propose and evaluate De-SAG in Section 5. The paper is concluded in Section 6 along with future work discussion.

2 Related Work

2.1 De-anonymization Attacks

Structure-based de-anonymization was first introduced by Backstrom et al. in [1], where they proposed both active and passive attacks. The primary idea of their attacks is to crate a subgraph and a connection pattern from the subgraph to the target users before the data is released. Then, after the data is released, the target users are de-anonymized by identifying the previously created subgraph and the connection pattern. However, the attacks in [1] are not scalable or tolerable to a change in graph topology during the data release process, i.e., it is not robust and thus can be easily defend against by obfuscating the graph structure. In [2], Narayanan and Shmatikov proposed a scalable and robust de-anonymization attack to graph data. That attack consists of two phases. In the first phase, a set of seed mappings from the anonymized graph to the auxiliary graph is identified. In the second phase, the de-anonymization is propagated from the seed mappings to the other nodes (users) in the anonymized graph leveraging several de-anonymization heuristics, e.g., eccentricity, node degrees, revisiting nodes, and reverse match.

In [3], Srivatsa and Hicks extended the structure-based de-anonymization technique to de-anonymize mobility traces. To achieve this, a contact graph is first constructed based on a mobility trace. Subsequently, a social graph is employed to de-anonymize the target contact graph. To perform the de-anonymization, three two-phase schemes are proposed. Similar to the attack in [2], the first phase of the three attacks is for seed identification. The second phase of the three attacks is based on three de-anonymization heuristics, namely Distance Vector (DV), Randomized Spanning Trees (RST), and Recursive Sub-graph Matching (RSM), respectively. In [4][5], Ji et al. proposed an iteration-based seed-free de-anonymization attack. During each de-anonymization iteration, two candidate sets of users V_1 and V_2 are selected from the anonymized and auxiliary graphs, respectively. Then, the users in V_1 are de-anonymized to the users in V_2 by minimizing an error function, which indicates the edge difference caused by a mapping scheme.

In [6], Nilizadeh et al. proposed a community-based de-anonymization technique for graph data, which can be used to enhance existing seed-based attacks. In this technique, a community-level de-anonymization is first performed. Subsequently, within each de-anonymized community, the user-level de-anonymization is conducted using existing attacks, e.g., [2]. In [7][8], Ji et al. designed a two-phase de-anonymization framework by considering more de-anonymization metrics, e.g., structural similarity, relative distance similarity, and inheritance similarity. They also addressed the scenario where the anonymized graph and the auxiliary graph have partial overlap.

Yartseva and Grossglauser proposed another two-phase de-anonymization attack in [15]. In the first phase of the attack, a set of seed mappings is identified. Then, in the second phase, all the (un-de-anonymized) neighbors of the de-anonymized users (including seed users) are considered as de-anonymization candidates and the pair of users who have the largest number of common de-anonymized neighbors are de-anonymized. Another similar two-phase attack is proposed in [16] by Korula and Lattanzi. After identifying the seeds in the first phase, the de-anonymization is propagated to the neighbors of the de-anonymized users. The pairs of users with the number of common de-anonymized neighbors greater than a threshold value will be de-anonymized. In [13], Chiasserini et al. studied the graph de-anonymization problem under the scale-free user relation model, which is considered to be more realistic. They again employed a two-phase de-anonymization framework.

In [9], Zhang et al. presented a structure-based de-anonymization attack to heterogeneous information networks, which are defined as graphs carrying heterogeneous (or multiple) relationships. They showed that how the extra information derived from heterogeneity can be used to improve the de-anonymization performance.

In addition to the above structure-based de-anonymization attacks, Wondracek et al. in [10] introduced a group membership information based attack to de-anonymize social network users. To conduct the attack, users' group membership information is first obtained leveraging a browser history stealing attack. Subsequently, the group membership information is used to distinguish (de-anonymize) different users. To some extent, the group membership information can be considered as one kind of attribute information. However, in addition to the membership information, other attribute information can also be utilized to conduct de-anonymization. Furthermore, structure-based de-anonymization is also powerful as shown in [1]-[7][15][16].

2.2 Defense

Generally, existing popular graph anonymization techniques can classified into four categories: k-anonymization based schemes [23][24][25][26], aggregation/class-based schemes [27][28][12], differential privacy based schemes [29][30], and random walk based schemes [31][32].

For k-anonymization based schemes, the basic idea is to make each node indistinguishable with at least $k-1$ other nodes with respect to some characteristic function (e.g., node degree) in the anonymized graph. Following this idea, Zhou and Pei proposed k-neighborhood [23], Liu and Terzi proposed k-degree [24], Zou et al. proposed k-automorphism [25], and Cheng et al. proposed k-isomorphism [26] to defend against structure-based graph de-anonymization attacks.

Aggregation/class-based anonymization schemes follow a similar idea as that in k-anonymization, where nodes are first grouped into classes and by topological operations, all the nodes within the same class are indistinguishable with respect to some defined characteristic functions [27][28][12].

Differential privacy was first proposed for statistical database query with strong privacy guarantee. Recently, many efforts have been spent to extend differential privacy to the scenario where data items have correlations, e.g., graph data. In [29], Sala et al. proposed a method to share graphs using differentially private graph models. In [30], Xiao et al. presented a data sanitization solution that infers a network's structure in a differentially private manner.

In [31], Mittal et al. proposed a random walk based graph anonymization scheme, under which each edge in the original graph is replaced by an edge generated by a random walk of length t. Later, this scheme was improved by Liu et al. in [32]. Instead of generating an edge via a fixed length random walk, Liu et al. proposed an adaptive random walk based graph anonymization method, where the random walk length is learned based on the local structural characteristics.

In [35][36], Ji et al. studied both the utility and the security performance of existing graph anonymization techniques. Through extensive analysis and empirical results, they demonstrated that existing anonymization techniques are still vulnerable to modern graph de-anonymization attacks. Furthermore, most existing defense techniques only focused on anonymizing graph structural information. It is seldom to see the anonymization scheme that accounts for both graph structural information and the attribute information.

2.3 De-anonymizability Analysis

Recently, in addition to studying the de-anonymization attacks, the issue of quantifying the de-anonymizability of graph data has also drawn much attention. In [14], Pedarsani and Grossglauser studied the de-anonymizability of graph data under the Erdős-Rényi (ER) model $G(n, p)$, where a graph consists of n users and any two users are connected (i.e., having a link) with probability p. They derived the structure conditions on the anonymized and auxiliary graphs to achieve perfect de-anonymization, i.e., successfully de-anonymize all the users in the anonymized graph. In [15], Yartseva and Grossglauser studied the seed-based de-anonymizability of graph data under the ER model. They specified the de-anonymization percolation condition of graph data (if the anonymized graph is percolated with respect to de-anonymization, the majority of the anonymized users can be successfully de-anonymized by structure-based de-anonymization attacks). However, a graph under the ER model has a Poisson degree distribution [4], which is different from most, if not all, of the real world graph data. Hence, the quantification under the ER model might not be applied to real world graph data. Similar to [15], Korula and Lattanzi studied the seed-based de-anonymizability of graph data under both the ER model and the Preferential Attachement (PA) model. They also specified the structure conditions for de-anonymizing a graph. However, the quantification in [16] is based on a strict assumption of existing dense seeds, which might be impractical.

In [4], Ji et al. quantified the structure-based de-anonymizability for graph data under the Configuration Model. They derived the structure conditions on the anonymized and auxiliary graphs for both perfect de-anonymization and partial de-anonymization (de-anonymizing a portion of anonymized users). In [6], Nilizadeh et al. analyzed the impacts of the community property on graphs' anonymity. They experimentally showed that the community information can be used to improve existing de-anonymization attacks. In [17][18], Ji et al. studied the seed-based de-anonymizability of graph data under both the ER model and a statistical model. Similar to [4], they specified the structure conditions for both perfect and partial de-anonymization.

2.4 Attack and Defense on SAG Data

In [11], Qian et al. proposed to de-anonymize social graphs and infer private attributes leveraging knowledge graphs, which carry both graph structural information and semantic information. In [33], Gong et al. adapted several representative supervised and unsupervised link prediction algorithms to SAG data and demonstrated the performance improvement for each algorithm with respect to both link prediction and attribute inference. They also evaluated the proposed

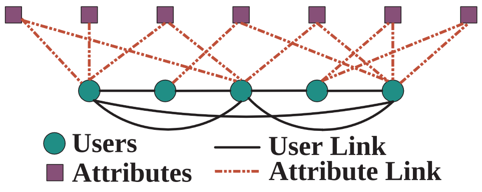


Fig. 1. The SAG model.

algorithms using Google+ datasets. Note that, we have a different focus than that in [33]. In this paper, we focus on node (user) privacy (node de-anonymization) instead of link or attribute privacy (link prediction and attribute inference). In [34], Jorgensen et al. proposed a method for publishing SAG data with formal privacy guarantees. They adapted existing graph models and introduced a new one, and then showed how to augment them with differential privacy. The output is a synthetic graph. Again, they focused on link (relationship) privacy and attribute privacy, while the target in this paper is to de-anonymize graph users, i.e., we focus on studying how to break node privacy.

2.5 Remark

In reality, most of the graph data are shared with the attribution information (non-PII) for data mining tasks or research purposes [20]. Therefore, it is meaningful to study new de-anonymization attacks leveraging structure and attribute information. More importantly, all the existing de-anonymizability analysis for graph data are structure-based. Fundamentally understanding the impacts of user-associated attributes on the anonymity of graph data is still an open problem. In this paper, to address this open problem, we first theoretically analyze the impacts of attributes on graph data's anonymity. Subsequently, we present a new de-anonymization framework, namely De-SAG, to de-anonymize SAG data leveraging both the structure and the attribute information. In summary, our attribute-based anonymity analysis together with existing structure-based de-anonymizability quantification provide data owners and researchers a more complete understanding of the privacy vulnerability of graph data. Further, the proposed technique (De-SAG) extends existing structure based de-anonymization attacks and improves the de-anonymization performance (as shown in the experiments).

3 Data Model, Preliminaries, and Definitions

To make the paper more readable, we summarize the notations in Table 1.

TABLE 1
Notations.

Notation	Description
G, G', and G''	original, anonymized, and auxiliary graphs
V, V', and V''	user (node) set
E, E', and E''	user to user link set
A, A', and A''	attribute set
W, W', and W''	user to attribute link set
l_{ij}	a link between users i and j
a_{ij}	a link between user i and attribute j
$\mathcal{A}_i$	the set of attributes associated with user i
n, N	the number of users, attributes
m, M	the number of user-user links, user-attribute links
(i, j)	user i is mapped (de-anonymized) to user j
π	a mapping from V' to V''
p_{ij}^{π}	the probability that i is mapped to j under π
$\mathbf{P}_i^{\pi}$	the mapping distribution of i under π
$H^{\pi}(i)$	the entropy of i under π
$H^{\pi}(G')$	the entropy of G' under π
$\mathbb{A}(G')$	the anonymity of G' under π
p	the existing probability of link a_{ij}
D_{ij}	the attribute difference of users i and j
μ_X	the mean value of binomial random variable X
q'	the probability that $a_{ij} \notin W$ and $a_{ij} \in W'$
q''	the probability that $a_{ij} \notin W$ and $a_{ij} \in W''$
κ	the density of user-attribute links

3.1 Data Model

Given a SAG, we model it as a graph $G = (V, E, A, W)$ as shown in Fig.1, where $V = \{i|i$ is a user$\}$ (the set of users), $E = \{l_{ij}|l_{ij}$ is a link between users i and $j\}$ (the set of all the links among users), $A = \{i|i$ is an attribute$\}$ (the set of all the non-PII associated with the users in V), and $W = \{a_{ij}|i \in V, j \in A, a_{ij}$ is a link between user i and attribute j, i.e., user i has attribute $j\}$ (the set of all the links between users and attributes). $\forall i \in V$, we denote the attributes associated with i by $\mathcal{A}_i$, i.e., $\mathcal{A}_i = \{j|j \in A, \exists a_{ij} \in W\}$. Furthermore, we define $n = |V|$ and $N = |A|$ to be the numbers of users and attributes, respectively.

3.2 De-anonymization

Given a raw SAG G, we assume that it will be anonymized before being shared/published. The anonymized G is denoted by $G' = (V', E', A', W')$ (we use an *apostrophe* to distinguish between the notations associated with G' from G when necessary). Note that, in G', although we cannot distinguish between the users in V' (we do not know the identities of the users in V'), we still know the attributes associated with each anonymized user since they are non-PII, e.g., in the published SAG data [19][20][21], the attributes (non-PII) associated with anonymized users are explicitly available. On the other hand, in reality, $\forall i \in V$, it is also possible that $\mathcal{A}_i \neq \mathcal{A}'_i$ after the anonymization process, i.e., the anonymization scheme may add some new attributes to and/or remove some existing attributes from a user.

For the adversaries, as in existing de-anonymization attacks [2][3][4][6], they try to de-anonymize G' leveraging some auxiliary graph denoted by $G'' = (V'', E'', A'', W'')$ (we use *double-apostrophe* to distinguish between the notations associated with G'' from G' and G when necessary), e.g., an adversary can leverage a Flickr graph to deanonymize a Twitter graph [2]. In reality, the auxiliary graphs can be obtained through multiple means, e.g., online crawling, data aggregation, data mining tasks, third-party information collection, public data sharing [2][4].

Without loss of generality, we assume $V' = V'' = V$ (although we do not know the users in V') and $A' = A'' = A$. Note that, as in [4][14], this assumption does not limit the results of this paper. When $V' \neq V''$ (respectively, $A' \neq A''$), the analysis in this paper is valid on V'_{new} and V''_{new}

(respectively, A'_{new} and A''_{new}) which are defined as $V'_{new} = V''_{new} = V' \cup V''$ (respectively, $A'_{new} = A''_{new} = A' \cup A''$); and the algorithm proposed in this paper can still work directly.

According to G' and G'', a de-anonymization attack/scheme can mathematically be defined as a mapping from V' to V'' [2][4][6][17], denoted by

$$\pi = V' \to V'' = \{(i, \pi(i) = j) | i \in V', j \in V''\}. \qquad (1)$$

Then, for convenience of discussion, $\forall i \in V'$, a correct de-anonymization of i is denoted by mapping (i, i), i.e., the *identical mapping* corresponds to the correct de-anonymization.

3.3 Anonymity of G'

Entropy has been widely used to quantify the randomness (uncertainty) of a process/system. Similarly, it can also be employed to measure the anonymity of G' given G'' and π [6]. Let π be an arbitrary de-anonymization scheme (mapping) from V' to V''. $\forall i \in V'$ and $\forall j \in V''$, let p^{π}_{ij} be the probability of the event that *i is mapped to j under π*. Then, $\forall i \in V'$, we denote its *mapping distribution* under π as $\mathbf{P}^{\pi}_i =< p^{\pi}_{i1}, p^{\pi}_{i2}, \cdots, p^{\pi}_{in} >$. Hence, the *uncertainty* of i under π can be measured by the *entropy carried by the mapping distribution* $\mathbf{P}^{\pi}_i$, which is formally defined as

$$H^{\pi}(i) = -\sum_{j=1}^{n} p^{\pi}_{ij} \log p^{\pi}_{ij}. \qquad (2)$$

Then, we define the *entropy/uncertainty of* G' under π as

$$H^{\pi}(G') = \frac{1}{n}\sum_{i=1}^{n} H^{\pi}(i), \qquad (3)$$

which is the *average entropy* of all the users in G'.

Let $H_{\max}(i) = \max\{H^{\pi}(i)\}$ and $H_{\max}(G') = \max\{H^{\pi}(G')\}$, respectively. Evidently, for each $i \in V'$, $H^{\pi}(i)$ is maximized when i can be mapped to each user in V'' equiprobably, i.e., i is perfectly anonymized. Hence, we have $H_{\max}(i) = \log n$. Similarly, we have $H_{\max}(G') = \log n$ when every user in G' achieves its maximum entropy, i.e., G' is perfectly anonymized. Then, based on $H^{\pi}(G')$ and $H_{\max}(G')$, we define the *anonymity* of G' under π as

$$\mathbb{A}(G') = \frac{H^{\pi}(G')}{H_{\max}(G')}. \qquad (4)$$

From the definition, we have $\mathbb{A}(G') \in [0, 1]$, where a large value of $\mathbb{A}(G')$ implies a better anonymity of G'. Specifically, $\mathbb{A}(G') = 0$ implies all the users in G' can be successfully de-anonymized under π while $\mathbb{A}(G') = 1$ implies that G' achieves the perfect anonymity.

4 Anonymity Analysis: From the Attribute Perspective

As we discussed in Sections 1 and 2, the structure-based de-anonymizability analysis for graph data has been studied in [4][14][15][16][17]. However, understanding the impacts of attributes on the anonymity/de-anonymizability of graph data is still an open problem. Furthermore, no existing de-anonymization scheme employs both the graph structure and the associated attributes to de-anonymize graph data. In this section, we address the first open problem by measuring the impacts of attributes on SAG data's anonymity. To be mathematically tractable, we conduct the analysis under a preliminary model first. Then, we generalize the analysis to the more complicated practical scenarios.

4.1 Preliminary Analysis

First, we conduct *attribute-based anonymity analysis* for SAGs under a random *Attribute Attachment* (A^2) model: given a SAG G, we assume that $\forall i \in V$ *and* $\forall j \in A$, *the existing probability of link* a_{ij} *is* p, i.e., $\Pr(a_{ij} \in W | \forall i \in V, \forall j \in A) = p$. Furthermore, we assume W' and W'' are random subsets of W: for each user-attribute link in W, it appeares in W' and W'' with positive probabilities p' and p'', respectively, i.e., $\Pr(a_{ij} \in W' | a_{ij} \in W) = p'$ and $\Pr(a_{ij} \in W'' | a_{ij} \in W) = p''$.

To facilitate our analysis, we introduce the concept of *Attribute Difference* (AD) between the users in V' and V''. $\forall i \in V'$ and $\forall j \in V''$, their AD is defined as

$$D_{ij} = (\mathcal{A}'_i \cup \mathcal{A}''_j) \setminus (\mathcal{A}'_i \cap \mathcal{A}''_j). \qquad (5)$$

In addition, before conducting the analysis, we give a lemma which will be used.

Lemma 1. *[14] Let X and Y be two binomial random variables with means μ_X and μ_Y, respectively. Then, if $\mu_X > \mu_Y$, $\Pr(X \leq Y) \leq 2\exp(-\frac{(\mu_X - \mu_Y)^2}{8(\mu_X + \mu_Y)})$.*

Let $\alpha = pp'(1-p'') + pp''(1-p')$ and $\beta = pp'(1-pp'') + pp''(1-pp')$. Furthermore, let $\vartheta = \frac{(\beta-\alpha)^2}{8(\beta+\alpha)}$. Then, we have the following theorem which quantifies the *attribute-based anonymity loss* of G'.

Theorem 1. *Let t be a natural number and $t \in [1, n-1]$. Then, (i) if $\vartheta \geq \frac{2\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\mathbb{A}(G') = \frac{\log t}{\log n}$; and (ii) if $\vartheta \geq \frac{3\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\mathbb{A}(G') = 0$, i.e., G' lost all of the anonymity.*

Proof Sketch: (i) To prove this conclusion, we first analyze the entropy of $\forall i \in V'$. Suppose i is de-anonymized to $j \in V''$ under some de-anonymization scheme π (we will discuss how to determine π later), i.e., $\pi(i) = j$. Then, we analyzed the AD caused by mapping (i, j). On one hand, if $j = i$, an AD will be induced if i has one attribute in exactly one of V' and V''. It follows that the AD corresponding to mapping $(i, j = i)$ is $D_{ij} = D_{ii} \sim \mathbf{B}(N, \alpha)$, where $\mathbf{B}(N, \alpha)$ is a *binomial variable* with parameters N and α. On the other hand, if $j \neq i$, the AD corresponding to mapping (i, j) is $D_{ij} \sim \mathbf{B}(N, \beta)$. Clearly, $\beta > \alpha$.

Let $\mathbf{E}$ be the event that *$\exists j \neq i$ such that $D_{ii} \geq D_{ij}$*. Then, according to Lemma 1, we have $\Pr(\mathbf{E}) \leq 2\exp(-\frac{(N\beta - N\alpha)^2}{8(N\beta + N\alpha)}) = 2\exp(-N\vartheta)$. Furthermore, the possible number of such events can be counted by t. Let $\mathbf{E}_t$ be the event that $\mathbf{E}$ *happens t times*. Then, we have $\Pr(\mathbf{E}_t) = C(n-1, t) \cdot \Pr(\mathbf{E})^t \cdot (1 - \Pr(\mathbf{E}))^{n-t} \leq C(n-1, t) \cdot \Pr(\mathbf{E})^t \leq \frac{(n-1)^t}{t!} \cdot 2\exp(-N\vartheta t) = \exp(t\ln(n-1) - \ln t!) \cdot 2\exp(-N\vartheta t) = 2\exp(t\ln(n-1) - \ln t! - N\vartheta t) \leq 2\exp(-2\ln n - 1) \leq \frac{1}{n^2}$. According to the Borel-Cantelli Lemma, we have $\Pr(\mathbf{E}_t) \to 0$ as $n \to \infty$. Therefore, when $\vartheta \geq \frac{2\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, with probability 1, $\mathbf{E}$ *happens less than t times*.

Algorithm 1: An implementation of π

```
for $i \in V'$ do
    sorting the users in $V''$ in the increasing order of $D_{ij}$ for $j \in V''$ and the sorted sequence is denoted as $< j_1, j_2, \cdots, j_n >$;
    mapping $i$ to $j_k$ $(1 \leq k \leq t)$ with probability $\frac{1}{t}$;
```

Based on the analysis, we define a simple de-aonymization scheme π as shown in Algorithm 1. From Algorithm 1, we have $\mathbf{P}_i^\pi =< p_{j_1}^\pi, p_{j_2}^\pi, \cdots, p_{j_t}^\pi, p_{j_{t+1}}^\pi, \cdots, p_{j_n}^\pi > =< \frac{1}{t}, \frac{1}{t}, \cdots, \frac{1}{t}, 0, \cdots, 0 >$. Furthermore, considering that $\Pr(\mathbf{E}_t) \rightarrow 0$, we conclude that π can successfully de-anonymize any user in V' with probability $\frac{1}{t}$. Then, $\forall i \in V'$, we have $H^\pi(i) = \log t$. It follows that $H^\pi(G') = \log t$ and thus $\mathbb{A}(G') = \frac{H^\pi(G')}{H_{\max}(G')} = \frac{\log t}{\log n}$.

(ii) Now, we prove the second conclusion. Let $\mathbf{E}_{all}$ be the event that *there exists some t such that $\mathbf{E}_t$ happens*. Then, $\Pr(\mathbf{E}_{all}) = \bigcup\limits_{t=1}^{n} \Pr(\mathbf{E}_t)$. Based on the *Boole's inequality*, we have $\Pr(\mathbf{E}_{all}) = \bigcup\limits_{t=1}^{n} \Pr(\mathbf{E}_t) \leq \sum\limits_{t=1}^{n-1} \Pr(\mathbf{E}_t) \leq \sum\limits_{t=1}^{n-1} 2\exp(t\ln(n-1) - \ln t! - N\vartheta t) = \sum\limits_{t=1}^{n-1} 2\exp(-3\ln n - 1) \leq \frac{1}{n^2}$. According to the Borel-Cantelli Lemma, we have $\Pr(\mathbf{E}_{all}) \rightarrow 0$ as $n \rightarrow \infty$, i.e., when $\vartheta \geq \frac{3\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\nexists t$ such that $\mathbf{E}_t$ happens. This further implies that with probability 1, $\forall i \in V'$ and $\forall j \in V''$, if $j \neq i$, $\Pr(D_{ii} < D_{ij}) \rightarrow 1$ as $n \rightarrow \infty$.

Algorithm 2: Another implementation of π

```
for $i \in V'$ do
    mapping $i$ to $j \in V''$ such that $j = \arg\min\limits_{j}\{D_{ij} | j \in V''\}$;
```

Based on our analysis, we give another simple implementation of π as shown in Algorithm 2. Under π, each user in V' can be successfully de-anonymized with probability 1 as $n \rightarrow \infty$. Therefore, $H^\pi(i) = 0$ $\forall i \in V'$. It follows $\mathbb{A}(G') = 0$, i.e., all the users can be successfully de-anonymized by π with probability 1. □

In Theorem 1, we analyzed the impacts of attributes (non-PII) on the anonymity/de-anonymizability of SAG data under the A^2 model. Based on our analysis, the attributes may also significantly reduce the anonymity of SAG data, which is similar to the graph structure (as shown in [4][14]-[17]). To make our analysis more practical, we extend it to general scenarios in the following subsection.

4.2 Extension: Practical Scenarios

In the analysis under the A^2 model, W' and W'' are two random subsets of W, which implies that $\forall i \in V$, $\mathcal{A}_i'$ and $\mathcal{A}_i''$ are two random subsets of $\mathcal{A}_i$. However, in reality, it is possible that some attributes in $\mathcal{A}_i$ may not appear in $\mathcal{A}_i'/\mathcal{A}_i''$ or some attributes in $A \setminus \mathcal{A}_i$ may appear in $\mathcal{A}_i'/\mathcal{A}_i''$. Therefore, in this subsection, we conduct the *attribute-based anonymity analysis* for SAG data under a more general model.

Under the general model, $\forall a_{ij} \in W$, a_{ij} appears in W' and W'' with probabilities p' and p'' respectively, i.e., $\Pr(a_{ij} \in W' | a_{ij} \in W) = p'$ and $\Pr(a_{ij} \in W'' | a_{ij} \in W) = p''$. Furthermore, $\forall a_{ij} \notin W$, it is appeared in W' and W'' with probabilities q' and q'' respectively, $\Pr(a_{ij} \in W' | a_{ij} \notin W) = q'$ and $\Pr(a_{ij} \in W'' | a_{ij} \notin W) = q''$. Let $W_U = \{a_{ij} | i \in V, j \in A\}$ be the *universal set* of all the possible user-attribute links. Then, $\forall a_{ij} \in W_U$, we have $\Pr(a_{ij} \in W | a_{ij} \in W_U) \underset{statistically}{\rightarrow} \frac{|W|}{|W_U|}$. Let $\kappa = \frac{|W|}{|W_U|}$ and define $\zeta = \kappa(p'(1-p'') + p''(1-p')) + (1-\kappa)(q'(1-q'') + q''(1-q'))$, $\delta = (\kappa p' + (1-\kappa)q')(\kappa(1-p'') + (1-\kappa)(1-q'')) + (\kappa(1-p') + (1-\kappa)(1-q'))(\kappa p'' + (1-\kappa)q'')$, and $\varpi = \frac{(\delta-\zeta)^2}{8(\delta+\zeta)}$. Then, we have the following theorem to quantify the impacts of attributes on the achievable anonymity of G'.

Theorem 2. *Let t be a natural number and $t \in [1, n-1]$. Then, (i) if $\delta > \zeta$ and $\varpi \geq \frac{2\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\mathbb{A}(G') = \frac{\log t}{\log n}$; and (ii) if $\delta > \zeta$ and $\varpi \geq \frac{3\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\mathbb{A}(G') = 0$.*

Proof: This theorem can be proven using similar techniques as in Theorem 1. □

In Theorem 2, we show the achievable anonymity of G' under a general statistical model. From Theorem 2, the condition on ϖ is similar to that of ϑ in Theorem 1. In addition, Theorem 2 has one more constraint $\delta > \zeta$, which actually comes from the fact that *for $i \in V$, the attributes that do not appeare in $\mathcal{A}_i$ may appear in $\mathcal{A}_i'$ and/or $\mathcal{A}_i''$*. Similar to Theorem 1, Theorem 2 also implies that the attributes associated with users (non-PII) may have significant impacts on the anonymity of SAG data.

4.3 Evaluation

In this subsection, we evaluate our *attribute-based anonymity analysis* both numerically and via experiments that leverage real world SAG datasets. Since there exists randomness in our evaluations, we repeat each group of evaluations 100 times. The results are the average of these 100 evaluations.

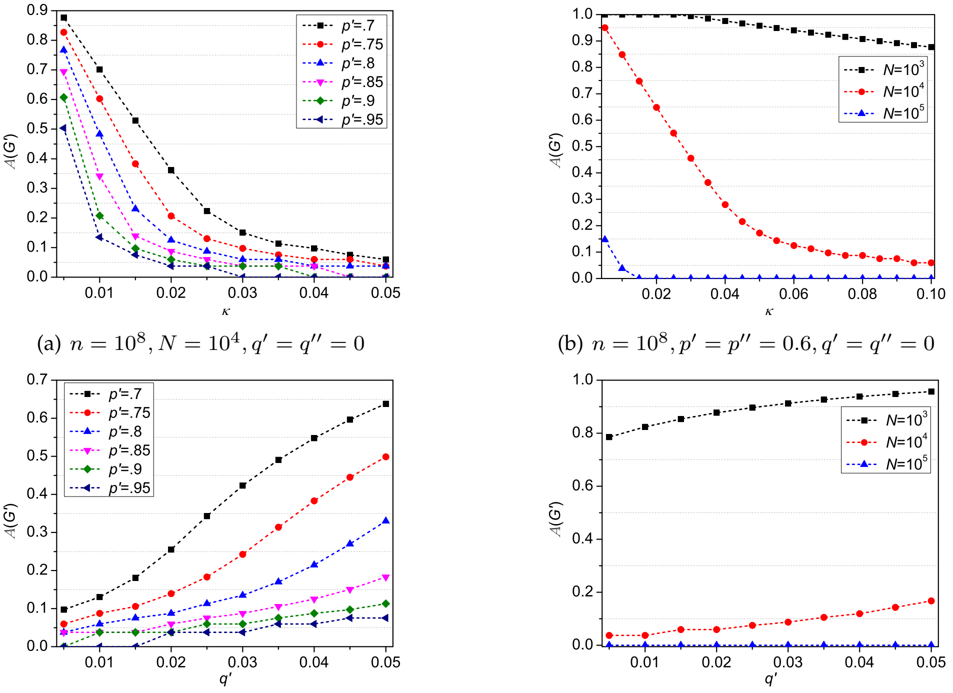
4.3.1 Numerical Evaluation

Since the analysis under the A^2 model can be viewed as a special case of that under the general model, our numerical evaluation follows the anonymity analysis under the general model, i.e., Theorem 2. Furthermore, to simplify the evaluation process, we set $p' = p''$ and $q' = q''$. Note that this setting does not limit our evaluation, and it can be removed directly by considering more scenarios.

In our evaluation, we first randomly generate a SAG G with the specified n, N, and κ. Subsequently, we generate G' and G'' from G according to p', p'', q', and q''. Finally, we evaluate $\mathbb{A}(G')$ based on Theorem 2. The detailed parameter settings are specified in each group of evaluations.

We show the evaluation results in Fig.2. We analyze the results as follows.

(1) From Fig.2 (a), with the increase of κ, $\mathbb{A}(G')$ decreases under different p'. This is because a larger κ implies more attribute information is associated with each user, statistically. Therefore, different users are more distinguishable with respect to attributes, i.e., with a higher probability $D_{ii} \leq D_{ij}$.



(a) $n = 10^8, N = 10^4, q' = q'' = 0$

(b) $n = 10^8, p' = p'' = 0.6, q' = q'' = 0$

(c) $n = 10^8, N = 10^4, \kappa = 0.05$

(d) $n = 10^8, p' = p'' = 0.8, \kappa = 0.06$

Fig. 2. Numerical evaluation of $\mathbb{A}(G')$.

TABLE 2
Data statistics.

	n	m	N	M	κ
GP1	4,693,129	47,130,325	991,545	3,644,103	7.83E-07
GP2	17,091,929	271,915,755	3,108,141	14,693,125	2.77E-07
GP3	26,244,659	410,445,770	4,147,389	19,344,382	1.78E-07
GP4	28,942,911	462,994,069	4,443,631	20,592,962	1.60E-07
GP5	107,614	13,673,453	19,044	387,261	1.89E-04
Facebook	4,039	88,234	1,283	37,257	7.19E-03
Twitter	81,306	1,768,149	216,839	1,245,234	7.06E-05

Furthermore, given κ, better anonymity is achieved when p' is smaller, e.g., given $\kappa = 0.02$, $\mathbb{A}(G') = 0.361$ when $p' = 0.7$ while $\mathbb{A}(G') = 0.038$ when $p' = 0.95$. This is because a larger p' implies more attributes can be preserved in G' and G'' (since $p'' = p'$), and thus it is more likely that $D_{ii} < D_{ij}$, i.e., a large p' implies more anonymity loss, which is consistent with our theoretical analysis.

(2) From Fig.2 (b), given n, p', p'', q', and q'', $\mathbb{A}(G')$ decreases when κ increases under different N. The reason is the same as in Fig.2 (a): a larger κ implies a higher probability of $D_{ii} < D_{ij}$, i.e., more anonymity loss. In addition, given κ, a larger N also implies more anonymity loss. For instance, given $\kappa = 0.065$, $\mathbb{A}(G') = 0.932$ when $N = 10^3$ while $\mathbb{A}(G') = 0.113$ when $N = 10^4$. This is because when κ is fixed, a larger N also implies richer attributes associated with each user and thus $D_{ii} < D_{ij}$ happens with a higher probability, i.e., $\mathbb{A}(G')$ decreases.

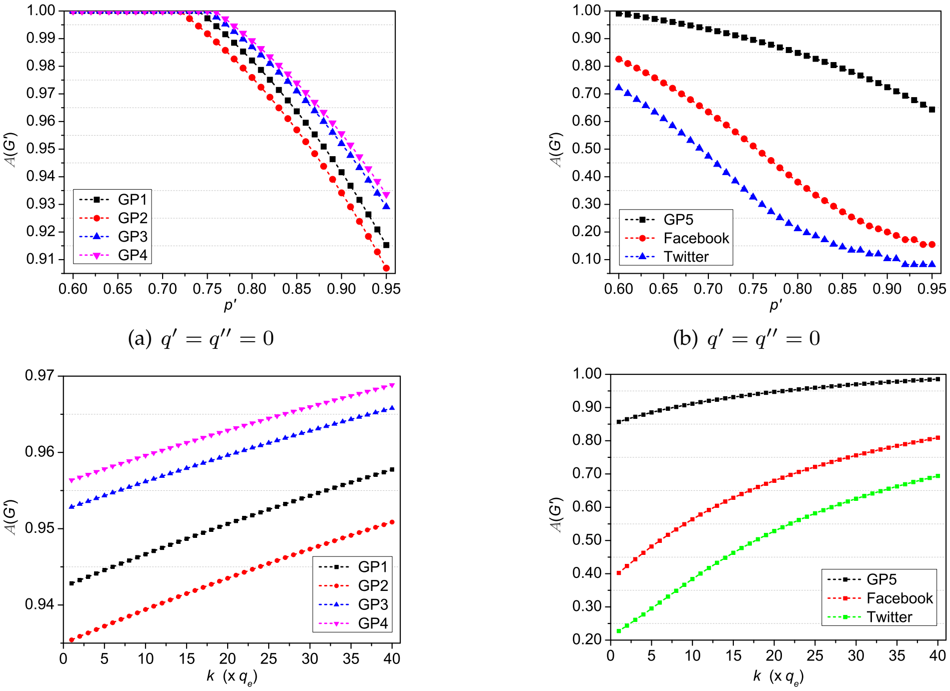
(3) Fig.2 (c) shows the impacts of q' (q'') on $\mathbb{A}(G')$. From Fig.2 (c), when q' increases, $\mathbb{A}(G')$ increases under different p'. This is because q' indicates the percentage of fake user-attribute relationships being added to G' and G'' ($q'' = q'$). A larger q' implies more link noise has been added to W' and W'' and thus a lower probability of $D_{ii} < D_{ij}$ has been induced, followed by the increase of $\mathbb{A}(G')$. Furthermore, given q', a smaller p' implies more anonymity can be achieved. The reason is the same as analyzed before: a smaller p' implies less common attributes are shared between G' and G''. Hence, better anonymity can be achieved by G'.

(4) In Fig.2 (d), we examine the impacts of q' and N on $\mathbb{A}(G')$. Again, when q' increases, $\mathbb{A}(G')$ also increases under different N. Additionally, given q', a larger N implies more anonymity loss. The reason is also the same as before. A larger N implies more attribute information is available for each user followed by less achievable anonymity according to our theoretical analysis.

4.3.2 Real World Data-based Evaluation

Now, we evaluate our attribute-based anonymity analysis leveraging real world SAG datasets.

Datasets. The employed SAG datasets include five Google Plus (GP) datasets, denoted by GPk ($1 \leq k \leq 5$) respectively [19][20][21], one Facebook dataset [21], and one Twitter dataset [21] as shown in Table 2, where $n = |V|$ (the number of users), $m = |E|$ (the number of user-user links), $N = |A|$ (the number of attributes), $M = |W|$



(a) $q' = q'' = 0$

(b) $q' = q'' = 0$

(c) $p' = p'' = 0.8$

(d) $p' = p'' = 0.8$

Fig. 3. Evaluation of $\mathbb{A}(G')$ leveraging on real data.

(the number of user-attribute links), and $\kappa = \frac{M}{n \cdot N}$ (the *connectivity* between users and attributes). All the SAG datasets include both the graph structure information and the attribute information (non-PII) associated with users. We introduce the datasets as follows.

GP is a social networking service launched in June 2011. It is designed to be a place to connect with friends and family. GP1, GP2, GP3, and GP4 are four GP datasets crawled in July 2011, August 2011, September 2011, and October 2011, respectively [19][20]. In addition to the social relationship information, there is also attribute information in the four GP datasets, e.g., gender, affiliation information, education, city. They are available under application. GP5 is another GP dataset which is publicly available at [21]. The attribute information in GP5 includes education, hometown, language, etc.

Facebook is one of the most popular social networking services in the world. It is designed as a social utility that connects people with friends and others who work, study, and live around them. The employed Facebook dataset is publicly available at [21]. The attribute information in Facebook includes birthday, education, location, employer, etc.

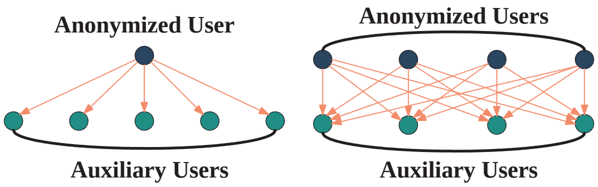
Twitter is also a popular online social networking service that enables users to send and read short messages named *tweets*. The employed Twitter dataset is publicly available at [21]. The attribute information in the Twitter dataset includes interests, cities, sports, websites, etc.

Results and Analysis. When conducting real world SAG data based evaluation, we first generate G' and G'' from each dataset according to the specified p', p'', q', and q''. Then, we evaluate the anonymity of G' following our analysis in Theorem 2. The evaluation results are shown in Fig.3, where the parameters are specified in each group of simulations. We analyze Fig.3 as follows:

(1) In Fig.3 (a) and (b), we show the impacts of p' on the achievable anonymity of each dataset, from which we can see that with the increase of p', the anonymity of each dataset decreases. For instance, when p' is increased from 0.6 to 0.8, $\mathbb{A}(\text{Facebook})$ is decreased from 0.826 to 0.381. The reason is similar to that in the numerical evaluation. A larger p' implies more attribute information is preserved in G' and G''. This further implies that the probability of $D_{ii} < D_{ij}$ increases, followed by the decrease of the anonymity. From Fig.3 (a) and (b), we can also see that the anonymity of GP1, GP2, GP3, and GP4 are higher than that of GP5, Facebook, and Twitter, e.g., when $p' = 0.85$, $\mathbb{A}(\text{GP1}) = 0.964$ while $\mathbb{A}(\text{Twitter}) = 0.145$. The main reason is that the connectivity of GPk $(1 \leq k \leq 4)$ is much smaller than that of GP5, Facebook, and Twitter. Therefore, GPk $(1 \leq k \leq 4)$ can achieve better anonymity, which is consistent with our analysis.

(2) Fig.3 (c) and (d) show the impacts of q', which is defined as $q' = k \cdot q_e = k \cdot \frac{(1-p')M}{|nN-M|}$ $(1 \leq k \leq 40)$ [1], on the anonymity of the seven datasets. From Fig.3 (c) and (d), the anonymity of each dataset increases with the increase of q', e.g., when q' is increased from $5q_e$ to $35q_e$, the anonymity of Facebook is increased from 0.482 to 0.785. The reason is that when q' increases, more fake user-attribute links (noise links) will be added to G' and G''. Then, the probability of $D_{ii} < D_{ij}$ is decreased, followed by the increase of the anonymity, which is consistent with our analysis.

[1] Here, we set q' in terms of p'. This is because we do not want to add too many fake user-attribute links in G' and G'' compared to the number of removed real user-attribute links. Otherwise, the data utility of G' for data mining tasks might be ruined.



(a) (b)

Fig. 4. User-based de-anonymization and set-based de-anonymization.

4.4 Discussion

From our summarization in Section 2, structure-based de-anonymizability analysis for graph data has been conducted in [4][14][16][15][17]. In this paper, for the first time, we study the impacts of attributes (non-PII) on the anonymity of graph data both theoretically and experimentally. Based on our analysis and evaluation results, we find that the attribute information associated with users may significantly reduce graph data anonymity. Therefore, our study together with existing structure-based de-anonymiability quantification research provides a much more complete understanding on the anonymity of graph data.

5 De-anonymization

In this section, we present a new de-anonymization framework, namely *De-SAG*, which considers both the graph structure and the attributes associated with users. Since *graph structure-based de-anonymization* has been well studied (as summarized in Section 2), we design De-SAG on top of existing structure-based de-anonymization attacks. Therefore, De-SAG can be considered as an enhanced version of existing de-anonymization attacks. To facilitate our design, we define a new notation $V''_{i,t}$ for $i \in V'$, which denotes the *t-most similar users of i in V'' with respect to attributes*. Evidently, $V''_{i,t}$ can be obtained using the same technique as in Algorithm 1 (i.e., let $V''_{i,t} = \{j_k | k = 1, 2, \cdots, t\}$). Let $D_{\max} = \max\{D_{ij} | i \in V', j \in V''\}$ be the maximum AD between any user in V' and any user in V''. Then, we define the *attribute similarity* of $i \in V'$ and $j \in V''$ as $1 - \frac{D_{ij}}{D_{\max}}$.

5.1 De-SAG

With respect to the *de-anonymization process*, existing structure-based de-anonymization attacks can be classified as *user-based de-anonymization schemes*, e.g., [2][6][15][16], and *set-based de-anonymization schemes*, e.g., [3][4][7] (the detailed explanations are given later). Since De-SAG is proposed on top of existing de-anonymization attacks, we present two implementations of the De-SAG framework based on the two classes of de-anonymization schemes.

5.1.1 User-based De-SAG

In user-based de-anonymization schemes [2][6][15][16], as shown in Fig.4 (a), during each de-anonymization iteration, one anonymized user i in V' is selected based on some criteria (e.g., having the maximum degree, having the most number neighbors being de-anonymized, having the most number of seed neighbors). Then, i is mapped (de-anonymized) to some user in V'' according to the proposed de-anonymization technique and the next de-anonymization iteration is started.

Algorithm 3: User-based De-SAG

```
while $V' \neq \emptyset$ do
    select $i$ from $V'$ as the user for de-anonymization
    according to the criteria in [2][6][15][16];
    map $i$ to some user $j$ in $V''_{i,t}$ according to the
    *enhanced* structure-based de-anonymization
    technique in [2][6][15][16], i.e., taking the *attribute
    similarity* as an extra mapping feature;
    $V' = V' \setminus \{i\}$;
    $V'' = V'' \setminus \{j\}$;
```

To enhance existing user-based de-anonymization attacks, we present an implementation of the user-based version of De-SAG as shown in Algorithm 3. From Algorithm 3, De-SAG basically follows the same process of existing user-based de-anonymization attacks. The primary improvements are (i) when de-anonymizing $i \in V'$, instead of considering all the users in V'' as candidates, we select the t-most-similar users of i with respect to attributes from V'' as candidate mappings. Here, t is a pre-defined parameter which controls the trade-off between de-anonymization accuracy and efficiency (a theoretically optimal t can be approximately estimated based on our anonymity analysis in Section 4[2]); and (ii) i is de-anonymized to one of the t-most similar users according to an *enhanced* version of existing structure-based de-anonymization attacks. In existing attacks, i is mapped to some user j in V'' according to the similarity of i and j's structural features, e.g., degree, betweenness centrality, closeness centrality [2][6][15][16]. In the enhanced version of existing attacks, De-SAG takes the attribute similarity as an extra mapping feature.

Let $O(T)$ and $O(S)$ be the time and space complexities of the enhanced user-based de-anonymization scheme in Algorithm 3, respectively. Then, the time complexity of De-SAG in Algorithm 3 is upper bounded by $O(n^2 \log n + T)$ since the candidate set size is reduced (the $O(n^2 \log n)$ time complexity is used to compute $V''_{i,t}$). The actual time complexity of De-SAG depends on the particular enhanced structure-based de-anonymization attack. The space complexity of De-SAG is also $O(S)$, i.e., De-SAG does not increase the space complexity of the enhanced scheme.

5.1.2 Set-based De-SAG

In set-based de-anonymization schemes [3][4][7], as shown in Fig.4 (b), during each de-anonymization iteration, a subset of un-de-anonymized users $\widetilde{V}'$ is selected from V' and a subset of auxiliary users $\widetilde{V}''$ is selected from V'',

[2] To estimate the theoretically optimal t, we first specify a temporary mapping from V' to V''. For instance, we can simply mapping V' to V'' according to the users' degree sequence: sorting the users in V' and V'' according to the degree non-increasing order and denoting the obtained user sequences as $< i_1, i_2, \cdots, i_n >$ and $< j_1, j_2, \cdots, j_n >$, respectively; and mapping i_k to j_k for $1 \leq k \leq n$. Second, we estimate G as $G = (V = V' = V'', E = E' \cup E'', A, W = W' \cup W'')$. Third, we estimate p, p', and p'' based on G, G', and G''. Fourth, we can estimate ϑ in terms of p, p', and p''. Finally, we estimate t as $t = \arg\min_t \vartheta \geq \frac{2 \ln n + t \ln(n-1) - \ln t! + 1}{Nt}$.

respectively. Subsequently, a *complete weighted bipartite graph* $\widetilde{G} = (\widetilde{V}, \widetilde{E})$ with $\widetilde{V} = \widetilde{V}' \cup \widetilde{V}''$ and $\widetilde{E} = \{l_{ij} | i \in \widetilde{V}', j \in \widetilde{V}''\}$ is constructed, where the weight of each link l_{ij}, denoted by $w(l_{ij})$, is determined according to the proposed de-anonymization techniques (usually, the weight of link l_{ij} measures how structurally similar i and j are and a larger weight means they are more structurally similar) [3][4][7]. After constructing $\widetilde{G}$, the de-anonymization problem reduces to a *Maximum Weighted Bipartite graph Matching problem* (MWBM) on $\widetilde{G}$. Finally, by addressing the MWBM problem on $\widetilde{G}$ (e.g., using the Hungarian algorithm), a mapping from $\widetilde{V}'$ to $\widetilde{V}''$ can be determined and the next de-anonymization iteration is started.

Algorithm 4: Set-based De-SAG

```
while $V' \neq \emptyset$ do
    determine $\widetilde{V}' \subseteq V'$ according to the criteria in [3][4][7];
    determine $\widetilde{V}'' \subseteq V''$ according to the criteria in [3][4][7];
    for $i \in \widetilde{V}'$ do
        $\widetilde{V}''_{i,t} \leftarrow V''_{i,t} \cap \widetilde{V}''$;
    construct a bipartite graph $\widetilde{G} = (\widetilde{V}' \cup \widetilde{V}'', \widetilde{E})$,
    where $\widetilde{E} = \{l_{ij} | i \in \widetilde{V}', j \in \widetilde{V}''_{i,t}\}$;
    for $l_{ij} \in \widetilde{E}$ do
        determine $w(l_{ij})$ according to the technique in [3][4][7];
        $w_a \leftarrow 1 - \frac{D_{ij}}{D_{\max}}$;
        $w(l_{ij}) \leftarrow c \cdot w(l_{ij}) + (1 - c) \cdot w_a$;
    de-anonymize $\widetilde{V}'$ based on $\widetilde{G}$ using the technique in [3][4][7];
    subtract the de-anonymized users from $V'$ and $V''$, respectively;
```

To enhance the set-based de-anonymization schemes [3][4][7], we present a set-based implementation of De-SAG as shown in Algorithm 4, where w_a denotes the *attribute similarity* of $i \in V'$ and $j \in \widetilde{V}''_{i,t}$, and $c \in [0, 1]$ is a pre-defined constant value. From Algorithm 4, De-SAG basically follows a similar process as existing set-based de-anonymization attacks. During each de-anonymization iteration, it improves existing schemes by further leveraging the attribute information. Specifically, De-SAG enhances existing set-based de-anonymization attacks in two perspectives. First, instead of constructing a *complete* bipartite graph, it reduces the number of links in $\widetilde{G}$ by setting $\widetilde{E} = \{l_{ij} | i \in \widetilde{V}', j \in \widetilde{V}''_{i,t}\}$. Second, it resets the weight associated with each link by taking account of the attribute similarity of two users (using $w(l_{ij}) \leftarrow c \cdot w(l_{ij}) + (1 - c) \cdot w_a$). Leveraging the two enhancements, (*i*) the computational complexity of existing set-based de-anonymization schemes can be reduced (since the mapping problem now is addressed on a non-complete bipartite graph); and (*ii*) the performance of existing set-based de-anonymization attacks can be improved (since the attribute similarity is used to enhance the de-anonymization process).

Let $O(T)$ and $O(S)$ be the time and space complexities of the enhanced set-based de-anonymization scheme in Algorithm 4, respectively. Then, similar to Algorithm 3, the time complexity of De-SAG in Algorithm 4 is upper bounded by $O(n^2 \log n + T)$ and the space complexity of De-SAG is also $O(S)$. Again, the actual time complexity of De-SAG depends on the particular enhanced structure-based de-anonymization attack.

5.2 Evaluation

In this subsection, we evaluate the performance of De-SAG and compare it with state-of-the-art de-anonymization attacks.

5.2.1 Evaluation Setting

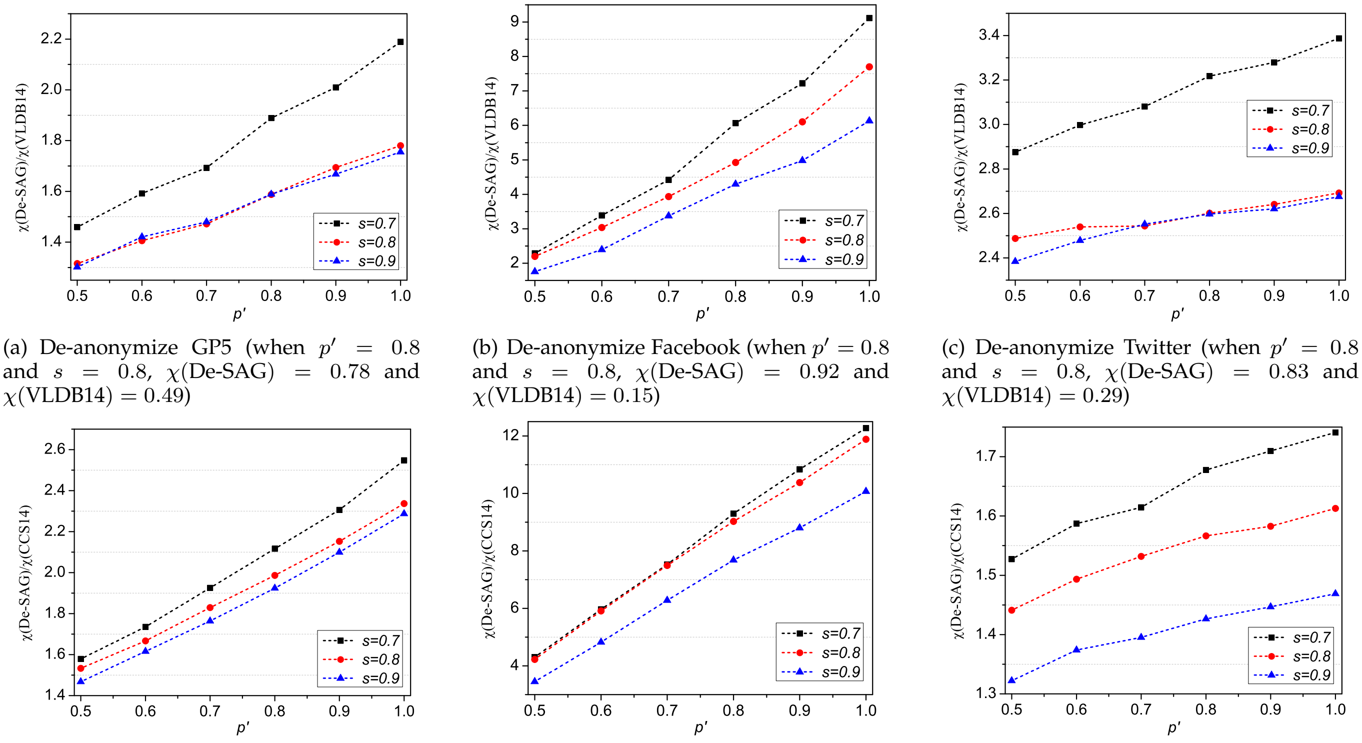
Since De-SAG has two implementations depending on the enhanced structure-based de-anonymization attacks, we compare De-SAG with the latest user-based de-anonymization scheme proposed in [16], denoted by VLDB14, and the latest set-based de-anonymization scheme proposed in [4], denoted by CCS14.

To conduct the evaluation, we employ three SAG datasets from Table 2: GP5, Facebook, and Twitter that are widely used in existing research [4][14][15][17][31], and follow the following methodology. First, given a raw dataset $G = (V, E, A, W)$ (i.e., GP5, Facebook, and Twitter here), we obtain the anonymized graph $G' = (V', E', A', W')$ and the auxiliary graph $G'' = (V'', E'', A'', W'')$ according to the parameter setting of each group of evaluations. When constructing (V', E') and (V'', E'') from G, i.e., the anonymized and auxiliary graphs, we employ the same technique as in [4][16] for fairness and accuracy. Specifically, we let $V' = V'' = V$ and E' and E'' are random subsets of E with each link in E appearing in E'/E'' with probability s, i.e., $\Pr(l_{ij} \in E' | l_{ij} \in E) = \Pr(l_{ij} \in E'' | l_{ij} \in E) = s$. For A' and A'', we assume $A' = A'' = A$ according to our model. We also determine W' and W'' according to our data model. Specifically, similar to the evaluation setting of our theoretical anonymity analysis, we set $\Pr(a_{ij} \in W' | a_{ij} \in W) = p' = \Pr(a_{ij} \in W'' | a_{ij} \in W) = p''$ and $\Pr(a_{ij} \in W' | a_{ij} \notin W) = q' = \Pr(a_{ij} \in W'' | a_{ij} \notin W) = q''$. Second, we employ VLDB14, CCS14, and De-SAG to de-anonymize G' leveraging G'', respectively. The *successful de-anonymization rate* of each de-anonymization algorithm is defined as $\chi(\cdot) = \frac{n_c}{n}$, where n_c is the number of users that have been successfully de-anonymized and $n = |V|$ is the total number of users in an anonyized dataset.

As summarized in Section 2, VLDB14 is a seed-based attack and CCS14 is a seed-free attack. Therefore, in our evaluation, we feed VLDB14 50 seed mappings, which are the top-50 users in G with respect to node degree. For other parameters, we specify them in each group of evaluations.

5.2.2 Results

In Fig.5, we show the impacts of p' on the performance of VLDB14, CCS14, and De-SAG when de-anonymizing GP5, Facebook, and Twitter. Specifically, we show the change of $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ with respect to the increase of p' in Fig.5 (a)-(c) and the change of $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ with respect to the increase of p' in Fig.5 (d)-(f), respectively. We analyze Fig.5 as follows.



(a) De-anonymize GP5 (when $p' = 0.8$ and $s = 0.8$, χ(De-SAG) $= 0.78$ and χ(VLDB14) $= 0.49$)

(b) De-anonymize Facebook (when $p' = 0.8$ and $s = 0.8$, χ(De-SAG) $= 0.92$ and χ(VLDB14) $= 0.15$)

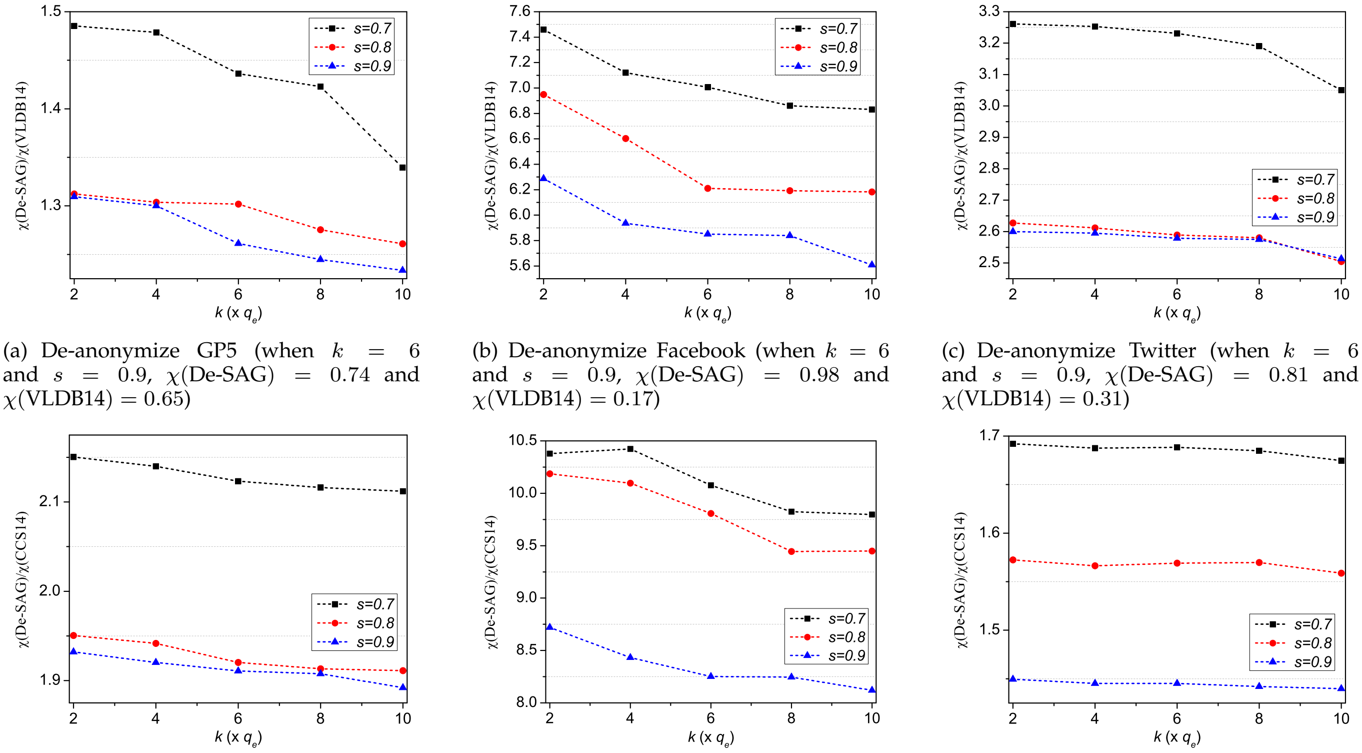
(c) De-anonymize Twitter (when $p' = 0.8$ and $s = 0.8$, χ(De-SAG) $= 0.83$ and χ(VLDB14) $= 0.29$)

(d) De-anonymize GP5 (when $p' = 0.8$ and $s = 0.8$, χ(De-SAG) $= 0.78$ and χ(VLDB14) $= 0.39$)

(e) De-anonymize Facebook (when $p' = 0.8$ and $s = 0.8$, χ(De-SAG) $= 0.92$ and χ(VLDB14) $= 0.11$)

(f) De-anonymize Twitter (when $p' = 0.8$ and $s = 0.8$, χ(De-SAG) $= 0.83$ and χ(VLDB14) $= 0.53$)

Fig. 5. De-SAG Evaluation (vs p'). Default setting: $q' = q'' = 0$ and $c = 0.5$.



(a) De-anonymize GP5 (when $k = 6$ and $s = 0.9$, χ(De-SAG) $= 0.74$ and χ(VLDB14) $= 0.65$)

(b) De-anonymize Facebook (when $k = 6$ and $s = 0.9$, χ(De-SAG) $= 0.98$ and χ(VLDB14) $= 0.17$)

(c) De-anonymize Twitter (when $k = 6$ and $s = 0.9$, χ(De-SAG) $= 0.81$ and χ(VLDB14) $= 0.31$)

(d) De-anonymize GP5 (when $k = 6$ and $s = 0.9$, χ(De-SAG) $= 0.74$ and χ(VLDB14) $= 0.4$)

(e) De-anonymize Facebook (when $k = 6$ and $s = 0.9$, χ(De-SAG) $= 0.98$ and χ(VLDB14) $= 0.12$)

(f) De-anonymize Twitter (when $k = 6$ and $s = 0.9$, χ(De-SAG) $= 0.81$ and χ(VLDB14) $= 0.57$)

Fig. 6. De-SAG evaluation (vs q'). Default setting: $p' = p'' = 0.8$ and $c = 0.5$.

(1) When p' increases, both $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ increase under different s. This is because when p' increases, more attribute information appears in both the anonymized graph and the auxiliary graph, i.e., the users in G' and G'' have more common attributes (which implies that the users in G' and G'' have better attribute similarity in Algorithms 3 and 4). Then, $\chi(\text{De-SAG})$ increases since more attribute information is available for de-anonymization, followed by the increase of $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$. Note that this result is also consistent with our theoretical analysis and experimental evaluation in Section 4: the increase of p' implies the decrease of the anonymity of G'.

(2) When de-anonymizing GP5, Facebook, and Twitter, on average, the successful de-anonymization rate of De-SAG is 1.63, 4.63, and 2.75 times of that of VLDB14 respectively, and is 1.94, 7.79, and 1.53 times of that of CCS14 respectively. This demonstrates that the attribute information is very powerful in enhancing existing structure-based de-anonymization attacks, which further confirms our attribute-based anonymity analysis (the attribute information can significantly reduce the anonymity SAG data).

(3) In most of the scenarios, De-SAG leads to more improvements compared to VLDB14 and CCS14 for smaller s than larger s. For instance, when de-anonymizing Facebook employing De-SAG and VLDB14 (Fig.5 (b)), on average, $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 5.42$ when $s = 0.7$, $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 4.65$ when $s = 0.8$, and $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 3.82$ when $s = 0.9$; and when de-anonymizing Twitter employing De-SAG and CCS14 (Fig.5 (f)), on average, $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 1.64$ when $s = 0.7$, $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 1.54$ when $s = 0.8$, and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 1.41$ when $s = 0.9$. This is because a small s implies that less links in E appeare in E' and E'', followed by less structural similarity between G' and G''. Therefore, the structure-based de-anonymization attacks VLDB14 and CCS14 will have a performance degradation. On the other hand, the attributes associated with users can provide relatively more useful information for successful de-anonymization.

Leveraging GP5, Facebook, and Twitter, we show the impacts of q' on the performance of VLDB14, CCS14, and De-SAG in Fig.6, where q' is defined as $q' = k \cdot q_e$ ($k = 2, \cdots, 10$, and $q_e = \frac{(1-p')M}{|nN-M|}$ which is the same as in Section 4). Specifically, the impacts of q' on $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ are shown in Fig.6 (a)-(c), and the impacts of q' on $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ are shown in Fig.6 (d)-(f), respectively. We analyze the results in Fig.6 as follows.

(1) When q' increases, both $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ decrease under different s. For instance, when de-anonymizing GP5 employing De-SAG and VLDB14 in the case of $s = 0.7$ (Fig.6 (a)), $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ is decreased from 1.49 to 1.34 when q' is increased from $2q_e$ to $10q_e$; and when de-anonymizaing GP5 employing De-SAG and CCS14 in the case of $s = 0.7$ (Fig.6 (d)), $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ is decreased from 2.15 to 2.11 when q' is increased from $2q_e$ to $10q_e$. This is because, as indicated in Section 4, with the increase of q', more fake user-attribute links will be added to G' and G'', and thus the benefit of employing the attribute information for de-anonymization is decreased, followed by the decrease of $\chi(\text{De-SAG})$. Then, both $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ decrease. This is consistent with our analysis and evaluation in Section 4.

(2) As in Fig.5, when de-anonymizing GP5, Facebook, and Twitter, on average, the successful de-anonymization rate of De-SAG is 1.33, 6.46, and 2.77 times of that of VLDB14 respectively, and is 1.99, 9.41, and 1.57 times of that of CCS14 respectively. This demonstrates that De-SAG can significantly improve existing structure-based de-anonymization attacks by taking account both the structure and the attribute information.

(3) Given q', similar to that in Fig.5, the improvements of De-SAG over VLDB14/CCS14 is higher for smaller s in most of the scenarios. For instance, when de-anonymizing Facebook (Fig.6 (b) and (e)), on average, $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 7.06$ when $s = 0.7$, $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 6.43$ when $s = 0.8$, and $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 5.9$ when $s = 0.9$; and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 10.1$ when $s = 0.7$, $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 9.8$ when $s = 0.8$, and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 8.33$ when $s = 0.9$. Again, this is due to the fact that the attributes associated with users can relatively provide more information for de-anonymization when less structural information is available.

5.3 Discussion

Based on our analysis and the evaluation results, De-SAG can significantly improve the performance of existing structure-based de-anonymization attacks by taking account both structure and attribute information. Therefore, in graph data sharing/publishing research, it is also important to protect the user-attribute relationships in addition to protecting the graph structure. However, to the best of our knowledge, most, if not all, of the existing graph anonymization techniques only consider to anonymize graph structure [22][29][31]. Hence, we plan to conduct SAG data anonymization research in the future by considering both the graph structure and the user-attribute relationships.

As shown in [35], even for structure-based de-anonymization attacks, it is difficult, if not impossible, to develop some effective anonymization techniques that can preserve all the data utility. In practice, SAG data can provide more potential auxiliary information for adversaries, which make the defense even more difficult. The possible countermeasures include: from the policy perspective, developing proper data access and publishing policies that can increase the difficulty of obtaining useful auxiliary information; from the technical perspective, one direction is that instead of trying to develop anonymization techniques that can preserve as much data utility as possible, we focus on designing application-aware anonymizaiton techniques with the target of preserving the desired data utility [35]. Another potential direction is to develop new privacy protection models, e.g., the adversarial learning based privacy protection model. Our attribute-based anonymity analysis and evaluation are expected to shed light on such research.

6 Conclusion and Future Work

In this paper, we study the impacts of the attribute information (non-PII) on the privacy of SAG data both theoretically and experimentally. First, we conduct an attribute-based anonymity analysis for SAG data. By careful quantification, we explicitly obtain the correlation between the graph anonymity and the associated attribute information. Through numerical and real world data-based evaluations, we validate our analysis and show that the attribute information may cause significant graph anonymity loss. Subsequently, according to our attribute-based anonymity analysis, we propose a novel de-anonymization framework, namely De-SAG, to graph data, which takes account both graph structure and attribute information. By extensive evaluation, we demonstrate that De-SAG can significantly improve the performance of state-of-the-art de-anonymization attacks. Our attribute-based anonymity analysis and de-anonymization framework are expected to fill the gap in understanding the actual privacy vulnerability of graph data and further shed light on future graph anonymization and de-anonymization research.

The future research directions of this paper are as follows. First, in addition to conducting structure-based and attribute-based anonymity analysis for graph data separately, we plan to analyze the privacy impacts of graph structure and attribute information simultaneously. Second, as we discussed in Section 5.3, it is expected to conduct SAG data anonymization research in the future by considering both the graph structure and the user-attribute relationships. Third, we did not consider the correlation that may exist among attributes. In practice, such correlation might be used for enhancing the capability of de-anonymizaiton attacks. Therefore, it is meaningful to extend our theoretical results and the de-anonymization framework to the scenario that accounts for such correlation.

Acknowledgment

We are very grateful to the anonymous reviewers for their time and valuable comments.

This work was partly supported by the Provincial Key Research and Development Program of Zhejiang, China under No. 2016C01G2010916, the Fundamental Research Funds for the Central Universities, the Alibaba-Zhejiang University Joint Research Institute for Frontier Technologies (A.Z.F.T.) under Program No. XT622017000118, and the CCF-Tencent Open Research Fund under No. AGR20160109.

References

[1] L. Backstrom, C. Dwork, and J. Kleinberg, *Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, WWW 2007.

[2] A. Narayanan and V. Shmatikov, *De-anonymizing Social Networks*, S&P 2009.

[3] M. Srivatsa and M. Hicks, *Deanonymizing Mobility Traces: Using Social Networks as a Side-Channel*, CCS 2012.

[4] S. Ji, W. Li, M. Srivatsa, and R. Beyah, *Structural Data De-anonymization: Quantification, Practice, and Implications*, CCS 2014.

[5] S. Ji, W. Li, M. Srivatsa, and R. Beyah, *Structural Data De-anonymization: Theory and Practice*, IEEE/ACM Transactions on Networking (ToN), Vol. 24, No. 6, pp. 3523-3536, 2016.

[6] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, *Community-enhanced De-anonymization of Online Social Networks*, CCS 2014.

[7] S. Ji, W. Li, M. Srivatsa, J. He, and R. Beyah, *Structure based Data De-anonymization of Social Networks and Mobility Traces*, ISC 2014.

[8] S. Ji, W. Li, M. Srivatsa, J. S. He, and R. Beyah, *General Graph Data De-Anonymization: From Mobility Traces to Social Networks*, ACM Transactions on Information and System Security (TISSEC), Vol. 18, No. 4, pp. 1-29, 2016.

[9] A. Zhang, X. Xie, K. Chen-Chuan, C. A. Gunter, J. Han, and X. Wang, *Privacy Risk in Anonymized Heterogeneous Information Networks*, EDBT 2014.

[10] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel, *A Practical Attack to De-Anonymize Social Network Users*, S&P 2010.

[11] J. Qian, X.-Y. Li, C. Zhang, and L. Chen, *De-anonymizing Social Networks and Inferring Private Attributes Using Knowledge Graphs*, INFOCOM 2016.

[12] C. Liu and P. Mittal *LinkMirage: Enabling Privacy-preserving Analytics on Social Relationships* NDSS 2016.

[13] C.-F. Chiasserini, M. Garetto, and E. Leonardi, *Social Network De-anonymization Under Scale-free User Relations*, IEEE/ACM Transactions on Networking (ToN), Vol. 24, No. 6, pp. 3756 C 3769, 2016.

[14] P. Pedarsani and M. Grossglauser, *On the Privacy of Anonymized Networks*, KDD 2011.

[15] L. Yartseva and M. Grossglauser, *On the Performance of Percolation Graph Matching*, COSN 2013.

[16] N. Korula and S. Lattanzi, *An Efficient Reconciliation Algorithm for Social Networks*, VLDB 2014.

[17] S. Ji, W. Li, N. Gong, P. Mittal, and R. Beyah, *On Your Social Network De-anonymizablity: Quantification and Large Scale Evaluation with Seed Knowledge*, NDSS 2015.

[18] S. Ji, W. Li, N. Gong, P. Mittal, and R. Beyah, *Seed based De-anonymizability Quantification of Social Networks*, IEEE Transactions on Information Forensics & Security (TIFS), Vol. 11, No. 7, pp. 1398-1411, 2016.

[19] The Google+ Dataset, http://www.cs.berkeley.edu/∼stevgong/dataset.html.

[20] N. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song, Evolution of Social-Attribute Networks: Measurements, Modeling, and Implications using Google+, IMC 2012.

[21] Stanford Large Network Dataset Collection, *http://snap.stanford.edu/data/index.html*.

[22] B. Zhou, J. Pei, and W.-S. Luk, *A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data*, KDD 2008.

[23] B. Zhou and J. Pei, *Preserving Privacy in Social Networks Against Neighborhood Attacks*, ICDE 2008.

[24] K. Liu and E. Terzi, *Towards Identity Anonymization on Graphs*, SIGMOD 2008.

[25] L. Zou, L. Chen, and M. T. Özsu, *K-Automorphism: A General Framework for Privacy Preserving Network Publication*, VLDB 2009.

[26] J. Cheng, A. Fu, and J. Liu, *K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks*, SIGMOD 2010.

[27] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, *Resisting Structural Re-identification in Anonymized Social Networks*, VLDB 2008.

[28] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava, *Class-based Graph Anonymization for Social Network Data*, VLDB 2009.

[29] A. Sala, X. Zhao, C. Wilson, H. Zheng, and B. Zhao, *Sharing Graphs using Differentially Private Graph Models*, IMC 2011.

[30] Q. Xiao, R. Chen, and K. Tan, *Differentially Private Network Data Release via Structural Inference*, KDD 2014.

[31] P. Mittal, C. Papamanthou, and D. Song, *Preserving Link Privacy in Social Network based Systems*, NDSS 2013.

[32] Y. Liu, S. Ji, and P. Mittal, *SmartWalk: Enhancing Social Network Security via Adaptive Random Walks*, ACM CCS 2016.

[33] N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. C. R. Shin, E. Stefanov, E. Shi, and D. Song, *Joint link prediction and attribute inference using a socialattribute network*, ACM Trans. Intell. Syst. Technol., Vol. 5, No. 2, 2014.

[34] Z. Jorgensen, T. Yu, and G. Cormode, Publishing Attributed Social Graphs with Formal Privacy Guarantees, SIGMOD 2016.

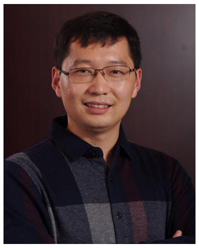
[35] S. Ji, P. Mittal, and R. Beyah, *Graph Data Anonymization, De-anonymization Attacks, and De-anonymizability Quantification: A Survey*, IEEE Communications Surveys & Tutorials (COMST), 2016.

[36] S. Ji, W. Li, P. Mittal, X. Hu, and R. Beyah, *SecGraph: A Uniform and Open-source Evaluation System for Graph Data Anonymization and De-anonymization*, USENIX Security 2015.

[37] https://www.facebook.com/

[38] https://twitter.com/

[39] https://plus.google.com/



Shouling Ji is a ZJU 100-Young Professor in the College of Computer Science and Technology at Zhejiang University and a Research Faculty in the School of Electrical and Computer Engineering at Georgia Institute of Technology. He received a Ph.D. in Electrical and Computer Engineering from Georgia Institute of Technology, a Ph.D. in Computer Science from Georgia State University. His current research interests include Big Data Security and Privacy, Big Data Driven Security and Privacy, and Adversarial Learning. He is a member of IEEE and ACM and was the Membership Chair of the IEEE Student Branch at Georgia State (2012-2013).

Ting Wang ia an Assistant Professor of Computer Science at Lehigh University. He is also affiliated with Data X, an interdisciplinary initiative that pushes the envelope of data analytics research. Prior to joining Lehigh, he was a Research Staff Member and security analytic leader at IBM Thomas J. Watson Research Center. He received a Ph.D. in Computer Science from Georgia Institute of Technology. His current research focuses on Computational Privacy, Cyber-Security Analytics and Network Science.

Jianhai Chen is currently a lecturer in the College of Computer Science and Technology at Zhejiang University. He received his B.S. degree in Applied Mathematics from Hunan University, and M.S. and PhD degrees in Computer Science and Technology from Zhejiang University. His research interests include Blockchain, Virtualization, and Cloud Computing. He is a member of IEEE and the ACM.

Weiqing Li is currently pursuing his M.S. degree in the School of Electrical and Computer Engineering at Georgia Institute of Technology. He received his B.S. degree from the School of Electrical and Computer Engineering at Georgia Institute of Technology. His research interests include Big Data Privacy and Network Security. He is a student member of ACM and IEEE.

Prateek Mittal is an assistant professor in the Department of Electrical Engineering at Princeton University. His research interests include the domains of privacy enhancing technologies, trustworthy social systems, and Internet/network security. His work has influenced the design of several widely used anonymity systems, and he is the recipient of several awards, including an ACM CCS outstanding paper. He served as the program co-chair for the HotPETs workshop in 2013 and 2014. Prior to joining Princeton University, he was a postdoctoral scholar at University of California, Berkeley. He obtained his Ph.D. in Electrical and Computer Engineering from University of Illinois at Urbana-Champaign in 2012

Raheem Beyah is the Motorola Foundation Professor and Associate Chair in the School of Electrical and Computer Engineering at Georgia Tech, where he leads the Communications Assurance and Performance Group (CAP) and is a member of the Communications Systems Center (CSC). Prior to returning to Georgia Tech, Dr. Beyah was an Assistant Professor in the Department of Computer Science at Georgia State University, a research faculty member with the Georgia Tech CSC, and a consultant in Andersen Consulting's (now Accenture) Network Solutions Group. He received his Bachelor of Science in Electrical Engineering from North Carolina A&T State University in 1998. He received his Masters and Ph.D. in Electrical and Computer Engineering from Georgia Tech in 1999 and 2003, respectively. His research interests include network security, wireless networks, network traffic characterization and performance, and critical infrastructure security. He received the National Science Foundation CAREER award in 2009 and was selected for DARPA's Computer Science Study Panel in 2010. He is a member of AAAS and ASEE, is a lifetime member of NSBE, and is a senior member of ACM and IEEE.