Quantifying Graph Anonymity, Utility, and De-anonymity

Shouling Ji[†], Tianyu Du[†], Zhen Hong[‡], Ting Wang[§], and Raheem Beyah[♯]
[†] College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China
[‡] Faculty of Mechanical Engineering & Automation, Zhejiang Sci-Tech University, Hangzhou, Zhejiang 310018, China
[§] Department of Computer Science and Engineering, Bethlehem, PA 18015, USA
[♯] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0765, USA
Email: {sji, zjradty}@zju.edu.cn, zhong@zstu.edu.cn, ting@cse.lehigh.edu, rbeyah@ece.gatech.edu

***Abstract*—In this paper, we study the correlation of graph data's anonymity, utility, and de-anonymity. Our main contributions include four perspectives. First, to the best of our knowledge, we conduct the *first* Anonymity-Utility-De-anonymity (AUD) correlation quantification for graph data and obtain *close-forms* for such correlation under both a preliminary mathematical model and a general data model. Second, we integrate our AUD quantification to SecGraph [31], a recently published Secure Graph data sharing/publishing system, and extend it to SecGraph+. Compared to SecGraph, SecGraph+ is an improved and enhanced *uniform* and *open-source* system for comprehensively studying graph anonymization, de-anonymization, and utility evaluation. Third, based on our AUD quantification, we evaluate the anonymity, utility, and de-anonymity of 12 real world graph datasets which are generated from various computer systems and services. The results show that the achievable anonymity/de-anonymity depends on multiple factors, e.g., the preserved data utility, the quality of the employed auxiliary data. Finally, we apply our AUD quantification to evaluate the performance of state-of-the-art anonymization and de-anonymization techniques. Interestingly, we find that there is still significant space to improve state-of-the-art de-anonymization attacks. We also explicitly and quantitatively demonstrate such possible improvement space.**

I. Introduction

Nowadays, many computer systems and services generate graph data [1], e.g., social network data [15], [23], [24], [25], mobility trace-based contact data [17], email network data [21], network topology data [24]. Mathematically, these data can be modeled by graphs, where nodes represent users and edges represent the relationships among users (e.g., friendships, contact-relationship, collaboration-relationship). Therefore, we refer to this type of data as *graph data* in this paper. Graph data are critical for academic research (e.g., the iDASH Privacy Protection Challenge [26], secure routing research [13]) and have many government, commercial, and healthcare applications (e.g., fraud detection [27], terrorist analysis [28], the Twitter-IBM project [29], disease propagation analysis [30]). Therefore, graph data are often shared or transferred to the research community, commercial partners, and other agencies for data mining tasks and analytics.

Meanwhile, since graph data usually carry sensitive information of users (e.g., the sexual contact data and other medical data [15]), privacy concerns arise when the data is shared/transferred. To protect users' privacy, many anonymization techniques have been proposed [1], e.g., the naive ID removal, random edge addition/deletion [2], k-anonymity-based techniques [3], [4], [5], [6], aggregation/cluster-based techniques [7], [8], Differential Privacy (DP)-based techniques [9], [10], [11], [12], and random walk-based technique [13]. Basically, the primary objective of those anonymization techniques is to protect users' privacy and simultaneously preserve as much data utility as possible. On the other hand, to break graph users' privacy, many structure-based de-anonymization attacks have also been presented [14]-[18], [21], which de-anonymize users according to their structural similarity in the anonymized and auxiliary graphs. In addition, to understand why graph data are vulnerable to structure-based de-anonymization attacks, several de-anonymization quantification techniques have been recently developed [19]-[22]. These techniques enable the understanding of the vulnerability of graph data by specifying the structural conditions for conducting perfect or partial de-anonymization.

Although several graph anonymization, de-anonymization, and de-anonymization quantification techniques have been proposed, there are still some important yet open problems in this area [1], such as: is there correlation between the anonymity, utility, and de-anonymity of graph data? if the correlation exists, how can it be quantified? what is the performance of state-of-the-art anonymization and de-anonymization techniques and is there room for improvement? Understanding these open problems are important for users and researchers. On one hand, it can help users and researchers understand how much utility is preserved after applying an anonymization scheme, what is the achievable data anonymity, what is the achievable de-anonymity given an auxiliary graph, as well as the correlation between anonymity, utility, and de-anonymity. On the other hand, it can also help users and researchers evaluate the performance of state-of-the-art anonymization and de-anonymization techniques compared to the achievable theoretical anonymity and de-anonymity.

Contributions. To address the aforementioned problems, we make the following contributions.

(i) We introduce three metrics to measure the anonymity, utility, and de-anonymity of anonymized graph data, respectively. Based on these metrics, we conduct a comprehensive quantification of the correlation of graph anonymity, utility, and de-anonymity under both the mathematical ER model and

a general data model. To the best of our knowledge, this is the first work on quantifying the AUD correlation of graph data and providing close-forms to explicitly demonstrate such correlation. Our results have important implications on graph data anonymization and de-anonymization research towards developing both powerful de-anonymization attacks and effective anonymization techniques.

(ii) Based on our AUD quantification, we implement SecGraph+ by adding a quantification module to SecGraph [31], a recently released uniform and open-source graph data anonymization and de-anonymization system. The extended SecGraph+ improves SecGraph from several perspectives: understanding the accurate AUD correlation of a graph dataset, guiding users/researchers to configure a proper anonymization algorithm, conducting objective-oriented on-demand evaluation and so on. SecGraph+ is a uniform and open-source system for advanced graph data anonymization, utility evaluation, and de-anonymization research.

(iii) Based on our correlation quantification, we conduct a large scale evaluation on the AUD of real world graph data leveraging 12 datasets that are generated from various computer systems and services. Our results demonstrate that the achievable anonymity/de-anonymity of graph data depends on multiple factors, e.g., the utility carried by the data, the quality of the employed auxiliary data.

(iv) Based on our AUD quantification, we evaluate the performance of state-of-the-art anonymization and de-anonymization techniques. Interestingly, we find that there is still significant room for state-of-the-art de-anonymization techniques to be improved. For instance, when using the latest seed-free de-anonymization attack ODA [21] to de-anonymize a Facebook dataset (64K users, 0.82M edges) that is anonymized by the state-of-the-art DP-based anonymization technique [9], [10], our evaluation shows that more than 83.4% theoretically de-anonymizable users cannot be de-anonymized by ODA.

II. Related Work

Attacks. Initially, in [14], Backstrom et al. introduced the structure-based de-anonymization attacks to graph data. Later, in [15], Narayanan and Shmatikov proposed a two-phase de-anonymization attack to de-anonymize large-scale graph data. Following [15], several improved two-phase de-anonymization attacks were introduced, e.g., the community-enhanced attack [16], mobility trace de-anonymization [17], adaptive de-anonymization [18], etc. In [21], Ji et al. proposed a seed-free de-anonymization attack on graph data by optimizing an objective function.

Defenses. In [2], Ying and Wu proposed two spectrum-preserved randomization techniques to anonymize graphs: *Rand Add/Del*, under which existing edges are randomly deleted and non-existing edges are randomly added, and *Rand Switch*, under which randomly selected pairs of edges are switched. To defend against neighborhood attacks, Zhou and Pei in [3] extended the k-anonymity to graph data. Another similar work is [4], where Liu and Terzi proposed a k-degree anonymous scheme for graph data. In [5], Zou et al. extended the k-anonymity idea to k-automorphism, under which a user cannot be distinguished from his/her $k-1$ symmetric users in an anonymized graph based on the structural information. Taking a similar idea, in [6], Cheng et al. proposed k-isomorphism, where the graph anonymization is achieved by forming k pairwise isomorphic subgraphs. In [7], Hay et al. proposed an aggregation-based approach to anonymize graph data. Similarly, in [8], Thompson and Yao presented a cluster-based graph anonymization technique.

To protect graph users' link information, Sala et al. introduced a graph anonymization technique using Differential Privacy (DP) [9]. Another similar work is [11], where Wang and Wu also developed a DP preserved technique for degree-correlation based graph anonymization. Based on the structural inference over the hierarchical random graph model, in [12], Xiao et al. also proposed a DP based graph anonymization scheme. In [13], Mittal et al. proposed a random walk based method to protect graph link privacy. Recently, Ji et al. implemented SecGraph, a uniform evaluation system for graph data anonymization and de-anonymization [31].

Theoretical Foundations. Recently, the de-anonymizability quantification problem has drawn a lot of attention from researchers. In [19], Pedarsani and Grossglauser studied the de-anonymizability of graph data under the ER model. In [20], Korula and Lattanzi quantified the de-anonymizability of graph data under both the ER model and the Preferential Attachment (PA) model. Ji et al. quantified the structural conditions of both perfect de-anonymization and partial de-anonymization under a general statistical data model for seed-free and seed-based attacks in [21] and [22].

III. System Model and Definitions

Utility. First, we model the raw data (e.g., social network data, email networks, contact graphs) for sharing/publishing by a graph $G^r = (V^r, E^r)$, where $V^r = \{1, 2, \cdots\}$ and $E^r = \{e_{i,j} | i, j \in V^r\}$ characterize the set of users and the set of relationships among users in the dataset respectively. Let $|V^r| = n$, i.e., the number of users is n. When sharing/publishing G^r, it is first anonymized by an arbitrary anonymization technique denoted by Π. Let $G^a = (V^a, E^a) = \Pi(G^r)$ be the anonymized graph. Without loss of generality, we assume $V^a = V^r$ (this is consistent with existing anonymization techniques [2]-[13]).

As shown in [2]-[12], the utility of G^a can be measured by many perspectives, e.g., degree distribution, cluster coefficient, network resilience etc. These metrics demonstrate the utility of the data from different perspectives. However, a general metric does not exist. On the other hand, we notice that existing utility metrics depend highly on how structurally/topologically similar G^r and G^a are. Therefore, we define an edge-based general utility metric μ. The objectives of defining μ are the following: consistent with existing utility metrics; sufficiently general to characterize the correlation between the raw and anonymized graphs; and mathematically tractable when quantifying the correlation of anonymity, utility, and de-anonymity. For $G^a = \Pi(G^r)$, it is defined as $\mu(G^a) = \frac{|E^a \cap E^r| + |\overline{E^a} \cap \overline{E^r}|}{|E^U|}$,

where $|\cdot|$ is the cardinality of a set, E^U is the universal set of all the possible edges that can be formed among users in V^r, $\overline{E^a} = E^U \setminus E^a = \{e_{i,j}|e_{i,j} \notin E^a\}$, and $\overline{E^r} = E^U \setminus E^r$. To be more accurate, we further define $\mu_1 = \frac{|E^a \cap E^r|}{|E^a|}$ and $\mu_0 = \frac{|\overline{E^a} \cap \overline{E^r}|}{|\overline{E^a}|}$. Then, $\mu(G^a) = \frac{\mu_1|E^a| + \mu_0|\overline{E^a}|}{|E^U|}$. Therefore, when $\mu_1 = \mu_0$, we have $\mu(G^a) = \mu_1 = \mu_0$. From the definition of $\mu(G^a)$, it measures the degree of G^a on preserving the structure (both the existing and the non-existing relationships) of G^r. We further experimentally demonstrate the performance of μ_Π in Section VII.

De-anonymity. To de-anonymize G^a, as in [15], [16], [17], [21], [22], we assume an auxiliary graph $G^u = (V^u, E^u)$ is available to the adversary. In reality, G^u can be obtained through multiple means, e.g., online crawling, data mining and aggregation, government publishing, third-party applications [15], [16], [17], [21]. Without loss of generality, we assume $V^u = V^r = V^a = V$ (this is a common assumption in existing analysis [19], [20], [21], [22]). When $V^u \neq V^a$, we can (i) either apply our analysis to the overlap users of V^a and V^u, or (ii) simply redefine $V^a = V^a \cup (V^u \setminus V^a)$ and $V^u = V^u \cup (V^a \setminus V^u)$ without changing E^a or E^u. Since G^u and G^r/G^a characterize the relationship of a same group of users, it is reasonable to assume G^u and G^r/G^a are correlated with each other. For instance, let G^r be an email network and G^u be an auxiliary Google+ graph of the same user set V. Then, for two users Alice and Bob in V, if they have a connection in G^r, they are also more likely to have a connection in Google+. To characterize this correlation between G^r and G^u, we statistically define $\Pr(e_{i,j} \in E^u | e_{i,j} \in E^r) = \tau$, and $\Pr(e_{i,j} \in E^u | e_{i,j} \notin E^r) = \gamma$, i.e., statistically, an edge appearers in G^u with probability τ when it also appears in G^r while with probability γ when it does not appear in G^r.

To be consistent with existing work [19], [20], [21], [22], we mathematically define a de-anonymization attack as a mapping $\sigma : V^a \to V^u$. Specifically, $\sigma := \{(i, \sigma(i))|i \in V^a, \sigma(i) \in V^u\}$. To simplify the discussion, a mapping $(i, \sigma(i))$ is correct when $i = \sigma(i)$ and incorrect otherwise. Given σ, let ω be the number of incorrect mappings in σ. Then, the ratio of successfully de-anonymized users in G^a under σ is defined as $\beta_\sigma = \frac{n-\omega}{n}$. Let $\mathbb{S}$ be the set of all the possible de-anonymization schemes. Since $|V^u| = |V^a| = n$, we have $n!$ possible mappings from V^a to V^u, i.e., $|\mathbb{S}| = n!$. The de-anonymity of G^a is defined as $\beta(G^a) = \max\{\beta_\sigma | \sigma \in \mathbb{S}\}$.

From the definition, the de-anonymity of G^a is measured by the maximum ratio of users that can be successfully de-anonymized. Intuitively, for an anonymized graph G^a, its practical de-anonymity depends on multiple factors, e.g., the correlation between the anonymized graph and the auxiliary graph. Therefore, it is difficult, if not possible, to derive the exact $\beta(G^a)$ for an arbitrary G^a. A practical quantification would seek to understand the de-anonymity of G^a relative to the utility carried by G^a. Toward this objective, we quantify the lower bound of the de-anonymity of G^a given the utility of G^a and an auxiliary graph in our AUD quantification.

Anonymity. We employ an information theoretical approach to define the anonymity of G^r/G^a given Π and σ, which is similar to that in [16]. For a user $i \in V^a$ and $\forall j \in V^u$, let $p_{i,j}$ be the probability of the event that i is mapped to (de-anonymized as) j in a de-anonymization scheme σ (i.e., $(i,j) \in \sigma$) and this mapping is a correct de-anonymization, i.e., $j = \sigma(i)$ (i and j correspond to the same user). For instance, if σ randomly and uniformly maps each $i \in V^a$ to any user $j \in V^u$, then we have $p_{i,j} = \frac{1}{n}$, i.e., i is successfully de-anonymized with probability $\frac{1}{n}$ under σ.

Based on the definition of $p_{i,j}$, we define $\mathbf{P}_{\Pi,\sigma}(i) = \{p_{i,j}|j \in V^u\}$ to be the mapping probability distribution of i under σ. Then, using information theory, the uncertainty of de-anonymizing i can be measured by entropy $H(i) = -\sum_{j \in V^u} p_{i,j} \log p_{i,j}$. When $\mathbf{P}_{\Pi,\sigma}(i) = \{p_{i,j} = \frac{1}{n} | j \in V^u\}$ (i is mapped to any user in V^u randomly and uniformly), i.e., the successful de-anonymization probability of i is $p_{i,j} = \frac{1}{n}$ for $\forall j \in V^u$ under Π, $H(i)$ reaches its maximum value $\log n$. In this scenario, an anonymization scheme Π is optimal from the perspective of protecting the privacy of i. On the other hand, if $\mathbf{P}_{\Pi,\sigma}(i) = \{p_{i,1} = 0, \cdots, p_{i,i-1} = 0, p_{i,i} = 1, p_{i,i+1} = 0, \cdots, p_{i,n} = 0\}$, i.e., the probability that i is successfully de-anonymized is 1, $H(i)$ reaches its minimum value 0. In this scenario, Π cannot protect the anonymity/privacy of i at all, i.e., the de-anonymization scheme σ can successfully break the privacy of i.

Based on $H(i)$, we can quantify the uncertainty of de-anonymizing G^a, denoted by $H(G^a)$, by the average entropy of all the users [16], i.e., $H(G^a) = \frac{1}{n} \sum_{i \in V^a} H(i)$. Let $H_{\max}(G^a)$ be the maximum entropy that G^a can achieve. Since $\max\{H(i)\} = \log n$, we have $H_{\max}(G^a) = \log n$. Here, if $H(G^a) = H_{\max}(G^a) = \log n$, G^a achieves the optimal anonymity. Then, the anonymity of G^a is defined as $\alpha(G^a) = \frac{H(G^a)}{H_{\max}(G^a)} = \frac{H(G^a)}{\log n}$, which measures how optimal G^a is on achieving uncertainty. Specifically, $\alpha(G^a) \in [0, 1]$, where 1 implies G^a achieves the best anonymity while 0 implies no anonymity at all.

From the anonymity definition, it is measured by the uncertainty of the process of de-anonymizing G^a. When studying the AUD correlation, our objective is to quantify the upper bound of the achievable $\alpha(G^a)$ relative to the utility preserved in G^a and the available auxiliary graph G^u.

In the remainder of this paper, we use $\mu = \mu(G^a)$, $\beta = \beta(G^a)$, and $\alpha = \alpha(G^a)$ for convenience of discussion. In addition, for the lowercase parameter x, we define $\overline{x} = 1 - x$, e.g., when $x = \mu$, $\overline{x} = \overline{\mu} = 1 - \mu$.

IV. AUD Quantification: ER Model

In this section, we quantify the AUD of graph data under the Erdős-Rényi (ER) model.

Preliminaries. Suppose G^r follows the ER model $G(n, p)$, i.e., there are n users in G^r and $\forall i, j \in V^r$, the edge $e_{i,j}$ appears in E^r with probability p ($\Pr(e_{i,j} \in E^r) = p$). When sharing/publishing G^r, it is first anonymized by an arbitrary anonymization scheme Π and the obtained anonymized graph is G^a. $\forall i \in V^a$, its neighborhood is defined as

$N_i^a = \{j | \exists e_{i,j} \in E^a\}$. Similarly, $\forall i \in V^u$, we define $N_i^u = \{j | \exists e_{i,j} \in E^u\}$.

Given $\sigma : V^a \rightarrow V^u$, to measure the quality of the mapping $(i,j) \in \sigma$, similarly as in [19], [21], [22], we define a Neighborhood Difference Function (NDF) $\Delta_{\sigma:(i,j)} = |(\bigcup_{v \in N_i^a} \{\sigma(v)\}) \setminus N_j^u| + |(\bigcup_{v \in N_j^u} \{\sigma^{-1}(v)\}) \setminus N_i^a|$, i.e., $\Delta_{\sigma:(i,j)}$ counts the neighborhood difference of $i \in V^a$ and $j \in V^u$ under σ. Then, to measure the performance of σ, we define the NDF of σ as $\Delta_\sigma = \sum_{i \in V^a} \Delta_{\sigma:(i,\sigma(i))}$.

Quantification. We quantify the AUD of an anonymized graph in this subsection. First, we quantify the NDF of a given σ. Given Π, G^a, G^u, and σ, let μ be the utility of G^a, $q_c(\mu) = p\mu_1\overline{\tau} + \overline{p} \cdot \overline{\mu_0} \cdot \overline{\gamma} + p\overline{\mu_1}\tau + \overline{p}\mu_0\gamma$, and $q_{i,c}(\mu) = (p\mu_1 + \overline{p} \cdot \overline{\mu_0})(p\overline{\tau} + \overline{p} \cdot \overline{\gamma}) + (p\overline{\mu_1} + \overline{p}\mu_0)(p\tau + \overline{p}\gamma)$. Then, we quantify the NDF of σ in Lemma 1. We omit the proof due to the space limitations.

Lemma 1. *If there are ω incorrect mappings in σ, $\Delta_\sigma \underset{n \rightarrow \infty}{\sim} B(\binom{n-\omega}{2}, q_c(\mu)) + B(\omega(n-\omega) + \binom{\omega}{2}, q_{i,c}(\mu))$, where $B(\cdot,\cdot)$ denotes a binomial variable.*

Based on Lemma 1, we can quantify the correlation of the utility, anonymity, and de-anonymity of an anonymized graph as shown in Theorem 1. The proof is omitted due to the space limitations.

Theorem 1. *Let $f(\mu) = \frac{(q_{i,c}(\mu) - q_c(\mu))^2}{8(q_{i,c}(\mu) + q_c(\mu))}$ be a utility function depending on the utility of G^a and ω be the number of possibly incorrectly de-anonymized users in a de-anonymization scheme. Then, when $q_{i,c}(\mu) > q_c(\mu)$ and $f(\mu) = \Omega(\frac{2\ln n + 1}{\omega n - \omega^2/2 - \omega/2})$[1], (i) $\beta = \Omega(\frac{n-\omega}{n})$; and (ii) $\alpha = O(\frac{\omega}{n}\log_n \omega)$.*

Remarks. In Theorem 1, we quantify the correlation between μ, β, and α. From the quantification results, the lower bound of the utility function $f(\mu)$ is defined by a decreasing function of ω (the number of possible incorrect mappings). When parameter ω increases, a looser condition is required for the utility function $f(\mu)$, followed by a lower de-anonymity β and a higher anonymity α of G^a are achievable. On the other hand, if higher de-anonymity is expected (i.e., lower anonymity can be achieved by G^a), a stricter condition is required on $f(\mu)$. Furthermore, from the proof of Theorem 1, When the specified conditions on $q_c(\mu)$, $q_{i,c}(\mu)$, and $f(\mu)$ are satisfied, a de-anonymization scheme σ that correctly de-anonymizes at least $n - \omega$ users can be found by a brute-force searching algorithm. Although the searching algorithm has a time complexity of $O(n!)$, which makes it computationally infeasible in reality, practical heuristics/approximation-optimization based de-anonymization attacks can be designed, e.g., [15], [17], [21]. Therefore, the significance of our quantification is to help users/researchers understand the theoretical correlation of anonymized graph data's utility, anonymity, and de-anonymity; and thus improve the anonymization/de-anonymization research.

[1]Here, $\Omega(\cdot)$ is employed to specify the lower bound in complexity theory. Formally, $f(x) = \Omega(g(x))$ implies $f(x) \geq k \cdot g(x)$ for some positive k, i.e., $\exists k > 0$, $\exists x > x_0$ such that $\forall x > x_0, f(x) \geq k \cdot g(x)$.

V. AUD Quantification: In General

In this section, we quantify the correlation of μ, β, and α under a general model, where the graph can have an arbitrary degree distribution.

We assume $G^r(V^r, E^r)$ can follow an arbitrary degree distribution. Let $m_r = |E^r|$. The graph density of G^r is defined as $\rho = \frac{m_r}{|E^U|} = \frac{2m_r}{n(n-1)}$. Let $\phi_c(\mu) = \mu_1\overline{\tau} + \overline{\mu_1}\tau$, $\phi_{i,c}(\mu) = \mu_1(\rho\overline{\tau} + \overline{\rho}\cdot\overline{\gamma}) + \overline{\mu_1}(\rho\tau + \overline{\rho}\gamma)$, $\psi_c(\mu) = \overline{\mu_0}\cdot\overline{\gamma} + \mu_0\gamma$, and $\psi_{i,c}(\mu) = \overline{\mu_0}(\rho\overline{\tau} + \overline{\rho}\cdot\overline{\gamma}) + \mu_0(\rho\tau + \overline{\rho}\gamma)$. Before quantifying the correlation of μ, β, and α, we first use Lemma 2 to quantify the NDF of a de-anonymization scheme σ, which has ω incorrect mappings. The proof is omitted due to the space limitations.

Lemma 2. *Let $\theta_{\min} = \min\{\phi_c(\mu), \psi_c(\mu)\}$, $\theta_{\max} = \max\{\phi_c(\mu), \psi_c(\mu)\}$, $\tau_{\min} = \min\{\phi_{i,c}(\mu), \psi_{i,c}(\mu)\}$, and $\tau_{\max} = \max\{\phi_{i,c}(\mu), \psi_{i,c}(\mu)\}$. If there are ω incorrect mappings in σ, $\Delta_\sigma \geq B(\binom{n-\omega}{2}, \theta_{\min}) + B(\omega(n-\omega) + \binom{\omega}{2}, \tau_{\min})$ and $\Delta_\sigma \leq B(\binom{n-\omega}{2}, \theta_{\max}) + B(\omega(n-\omega) + \binom{\omega}{2}, \tau_{\max})$.*

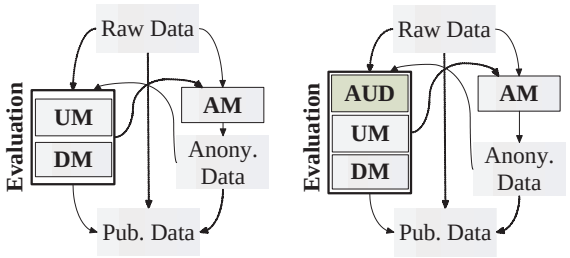
In Lemma 2, we derive the lower and upper bounds of Δ_σ for a given σ. Based on Lemma 2, we quantify the correlation of μ, β, and α under a general data model in Theorem 2. The proof is omitted due to the space limitations.

Theorem 2. *Let $g(\mu) = \frac{(\tau_{\min} - \theta_{\max})^2}{8(\tau_{\min} + \theta_{\max})}$ be a utility function depending μ, and ω be the number of possibly incorrectly de-anonymized users in a de-anonymization scheme. Then, when $\tau_{\min} > \theta_{\max}$ and $g(\mu) = \Omega(\frac{2\ln n + 1}{\omega n - \omega^2/2 - \omega/2})$, (i) $\beta = \Omega(\frac{n-\omega}{n})$; and (ii) $\alpha = O(\frac{\omega}{n}\log_n \omega)$.*

Remarks. From Theorem 2, the correlation of μ, β, and α under a general model is similar to that under the ER model. However, they are different with respect to required conditions and generality/applicability. Fundamentally, to achieve the same anonymity/de-anonymity, the conditions under the general model (specified by $g(\mu)$, $\tau_{\min}$, and $\theta_{\max}$) are stricter than that under the ER model (specified by $f(\mu)$, $q_c(\mu)$, and $q_{i,c}(\mu)$). On the other hand, the quantification in Theorem 1 is dedicated for graphs under the ER model while the quantification in Theorem 2 is applicable to graphs following any distribution.

VI. SecGraph+: An Enhanced Uniform and Open-source Secure Graph Data Sharing/Publishing System

SecGraph and Limitations. In [31], [32], Ji et al. developed a uniform and open-source Secure Graph data sharing and publishing system, SecGraph. The architecture of SecGraph is shown in Fig.1 (a). SecGraph consists of three main modules: the Anonymization Module (AM) where 11 state-of-the-art graph anonymization algorithms are implemented, the Utility Module (UM) where 12 graph utilities (e.g., degree, joint degree, cluster coefficient, path length) and 7 application utilities (e.g., influence maximization, community detection,



(a) SecGraph [31] (b) SecGraph+

Fig. 1. Architectures of SecGraph and SecGraph+. AM = Anonymization Module, UM = Utility Module, and DM = De-anonymization Module.

secure routing) are implemented, and the De-anonymization Module (DM) where 15 modern de-anonymization attacks are implemented. SecGraph has meaningful implications to both data owners and researchers. For data owners, they can use SecGraph to anonymize their data, measure the anonymized data's graph and application utilities, and evaluate the data's actual vulnerability against de-anonymization attacks. For researchers, they can use SecGraph to conduct fair analysis and evaluation of existing and newly developed anonymization and/or de-anonymization techniques.

Although SecGraph is meaningful to both data owners and researchers, it still has several limitations in helping users and researchers understand the precise anonymity, utility, and de-anonymity of anonymized graph data. First, it is difficult to employ SecGraph to understand the accurate correlation of the anonymity, utility, and de-anonymity of a graph dataset. As shown in [31] and Fig.1 (a), a user/researcher must repeatedly execute the "data - AM - anonymized data - UM/DM" process several times to obtain some intuition of the AUD correlation of a graph dataset. However, such AUD correlation intuition could be biased depending on the evaluation and parameter settings and thus misleading. Second, it is not trivial to employ SecGraph to conduct utility/anonymity-oriented on-demand evaluation. For instance, with the objective of preserving a specific amount of data utility, it is unclear what is the maximum achievable anonymity and how to configure an anonymization algorithm to achieve such anonymity. Again, a user must repeat the "data - AM - anonymized data - UM/DM" process to obtain some anonymized data with a higher anonymity. However, it is still unclear how this achieved anonymity compares to the maximum achievable anonymity. Furthermore, the evaluation process could be very time consuming and inefficient. Similarly, when evaluating the performance of a de-anonymization attack, it is difficult to tell how optimal this de-anonymization attack is compared to the achievable de-anonymity.

SecGraph+: Towards Comprehensive Graph Anonymization, Utility Evaluation, and De-anonymization. To address the limitations of SecGraph, we enhance it by integrating our AUD quantification and implement SecGraph+. Specifically, in SecGraph+, we add one more theoretical evaluation module, named the AUD correlation quantification module, as shown in Fig.1 (b). Since SecGraph is a module-based open-source system (available at [32]), the proposed SecGraph+ can be implemented directly by integrating our AUD correlation quantification techniques.

TABLE I
DATA STATISTICS.

Name	Type	n	m	ρ	d_{avg}
Wiki (WK)	Wiki	2.4M	5M	1.63E-6	3.9
Gnutella (GT)	P2P	36.7K	88.3K	1.32E-4	4.8
YouTube (YT)	SN	1.1M	3M	4.64E-6	5.3
Oregon (OG)	AS	11.5K	32.7K	4.98E-4	5.7
Brightkite (BK)	LSN	58K	.2M	1.32E-4	7.5
Gowalla (GW)	LSN	.2M	1M	4.92E-5	9.7
Enron (EN)	Email	36.7K	.2M	3.19E-4	10.7
Skitter (SK)	AS	1.7M	11.1M	7.73E-6	13.1
Facebook (FB)	SN	64K	.82M	4.02E-4	25.64
Google+ (G+)	SN	4.7M	90.8M	8.24E-6	38.7
Twitter (TW)	SN	.5M	14.9M	1.20E-4	54.8
Flickr (FL)	SN	80.5K	5.9M	1.82E-3	146.56

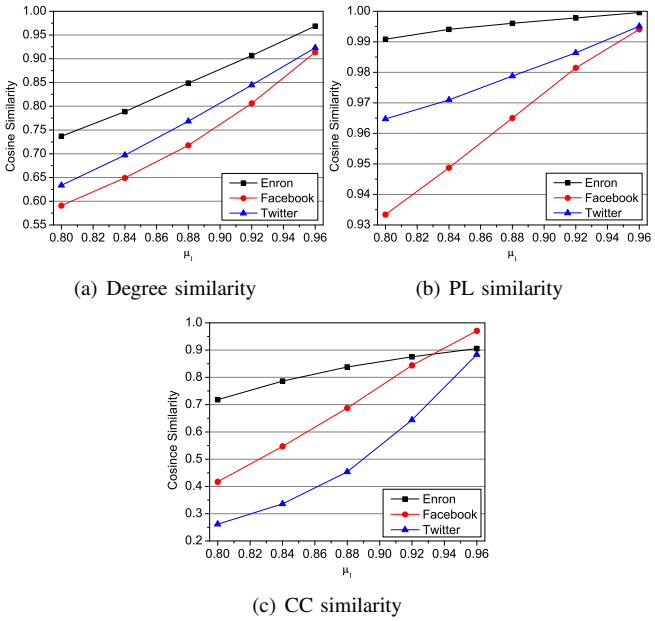
According to our evaluation in Section VII and Section VIII, the extended SecGraph+ is more useful and meaningful to users and researchers compared to SecGraph. First, instead of repeating the "data - AM - anonymized data - UM/DM" process to obtain some intuition, users and researchers can accurately evaluate the explicit AUD correlation of a graph dataset (as shown by our evaluation in Section VII). Second, the extended SecGraph+ is more helpful and more time efficient in conducting objective-oriented on-demand evaluation. For instance, after specifying the utility demand, users and researchers can employ SecGraph+ to statistically evaluate the upper bound of the achievable anonymity and the lower bound of the achievable de-anonymity. This can guide users/reserachers to further determine and configure a proper anonymization technique and/or a powerful de-anonymization attack (as shown in Section VII and Section VIII). Finally, SecGraph+ is also more meaningful for future anonymization, de-anonymization, and utility evaluation research. The quantification results of the AUD module can be used not only for evaluating the performance of existing anonymization/de-anonymization techniques, but it can also shed light on developing improved anonymization/de-anonymization techniques that (approximately) achieve the theoretical anonymity/de-anonymity bounds.

SecGraph+ is an advanced open and uniform system for graph data anonymization, de-anonymization, and utility evaluation (the SecGraph+ website is anonymized according to the conference's anonymous review rule). Integrating AUD to a publicly available research system/tool facilitates the application of our quantification techniques to future graph anonymization and de-anonymization research. Since SecGraph has been well evaluated in [31], we focus on evaluating our AUD quantification (the AUD module of SecGraph+) leveraging real world graph data in the following sections.

VII. UTILITY METRIC AND AUD EVALUATION

A. Datasets

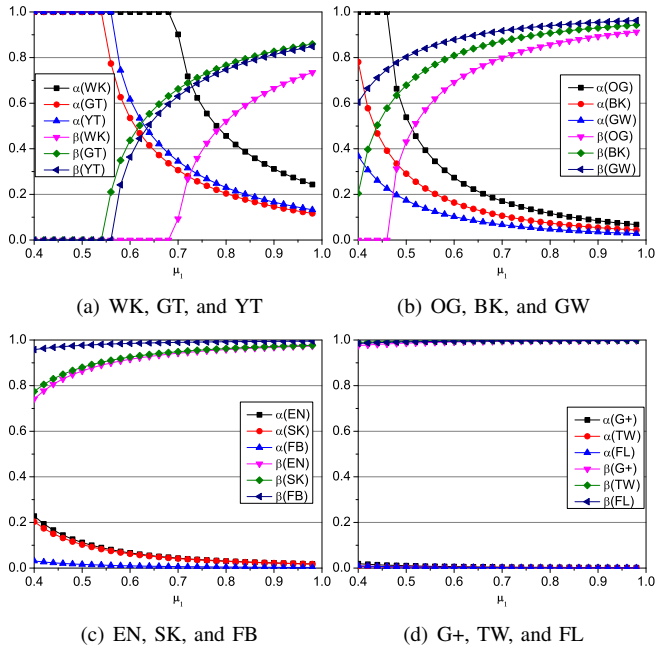
In the evaluation, we employ 12 real world graph datasets, which are generated from various computer systems: Social



(a) Degree similarity (b) PL similarity

(c) CC similarity

Fig. 2. The performance of the utility metric μ.



(a) WK, GT, and YT (b) OG, BK, and GW

(c) EN, SK, and FB (d) G+, TW, and FL

Fig. 3. AUD vs. μ_1.

Networks (SN), Location-based Social Networks (LSN), Email networks, WikiTalk networks (Wiki), P2P networks (P2P), and Autonomous Systems (AS). All the datasets are now publicly available and can be found at Berkeley Datasets [23], Stanford SNAP [24], and ASU Datasets [25]. We show the basic statistical information of the 12 datasets in Table I, where n, m, ρ, and d_{avg} denote the number of users, the number of edges, the graph density, and the average degree of each user, respectively.

B. Performance of the Utility Metric μ

Before evaluating our AUD quantification, we first examine the performance of our utility metric μ. According to the definition of μ, it measures the performance of G^a on preserving the structure (both the existing and the non-existing relationships) of G^r. Furthermore, since μ is defined based on μ_1 and μ_0, we examine the effectiveness of μ by evaluating the utility of G^a with respect to different μ_1 and μ_0. Due to the space limitations, here, we employ three datasets Enron, Facebook, and Twitter as example datasets for the evaluation, and the evaluated utilities of G^a are Degree distribution (Deg), Path Length distribution (PL), and Cluster Coefficient distribution (CC). The reason we choose to evaluate Deg, PL, and CC is because they are the most fundamental graph utilities and most of the other graph utilities (e.g., infectiousness, reliable email, secure routing, and influence propagation) are highly dependent on them [2]-[11].

Methodology. We employ the same utility evaluation methodology as in [31], [32]. Specifically, (i) given μ_1 and μ_0 and a raw dataset G^r, we employ the *Rand Add/Del* anonymization algorithm in [2] to anonymize G^r (by deleting existing edges and adding new edges) such that the obtained anonymized graph G^a has utility of μ_1 and μ_0; (ii) compute the Deg, PL, and CC utilities of both G^r and G^a; and (iii) compute the cosine similarity of each utility of G^r and G^a [2].

Results. We show the evaluation results in Fig.2, where when changing μ_1 in each evaluation, we set $\mu_0 = 1 - \frac{\mu_1 \cdot |E^a|}{|E^U \setminus E^a|}$ (note that, μ_0 is an increasing function of μ_1)[3]. From Fig.2, we have the following two observations.

First, with the increase of our utility metric μ_1/μ_0, all the three fundamental graph utilities Deg, PL, and CC are also increasing, which demonstrates that our utility metric is consistent with existing utility metrics. The reason is because μ_1 and μ_0 measures the degree of G^a to preserve the existing and non-existing relationships of G^u. When G^a and G^u share more common relationships, they are more structurally similar followed by high utility of G^a. Furthermore, based on Theorems 1 and 2, μ_1/μ_0 also makes our AUD quantification tractable. Therefore, our utility metric is effective.

Second, the changing magnitude of PL is smaller than the other two utilities with the increase of μ_1/μ_0. This is because the graph diameters of Enron, Facebook, and Twitter are 10, 10, and 7 respectively, which are relatively small. Therefore, the impact of anonymization (adding/deleting edges) to PL is also relatively small. On the other hand, when μ_1/μ_0 is small, a significant number of relationships in G^r have been changed in G^a. Since Deg and CC are local graph properties, they are more sensitive to local edge changes, i.e., μ_1/μ_0.

[2]Given two vectors, $\mathbf{A} =< A_1, A_2, \cdots, A_N >$ and $\mathbf{B} =< B_1, B_2, \cdots, B_N >$, the cosine similarity between $\mathbf{A}$ and $\mathbf{B}$ is defined as $\cos(\mathbf{A}, \mathbf{B}) = \frac{\sum_{i=1}^{N} A_i \times B_i}{\sqrt{\sum_{i=1}^{N}(A_i)^2} \times \sqrt{\sum_{i=1}^{N}(B_i)^2}}$. For our purpose, when evaluating the Deg, PL, and CC utilities of the anonymized graph, we use the cosine similarity of their distributions in the anonyimized graph and auxiliary graph.

[3]The purpose of this setting is to make G^a have relatively similar performance on preserving the existing and non-existing relationships of G^r.

C. AUD Evaluation

Methodology. For each dataset, it is the raw graph G^r in our evaluation. Then, given μ_1, μ_0, τ, and γ, the structures of both G^a and G^u can be derived from G^r. Finally, we apply our quantification technique in Section V to quantify the anonymity and de-anonymity of G^a based on G^u. Specifically, according to our proofs in Theorem 1 and Theorem 2, statistically, the optimum de-anonymization scheme (mapping) incudes the least NDF. Therefore, after specifying G^a and G^u, we can drive the anonymity and de-anonymity of G^a based on the utility preserved in G^a and G^u (relative to the raw data G^r) using Theorem 2 (the proof of Theorem 2). Note that, here, we are not trying to quantify the exact anonymity/de-anonymity of G^a (and thus, we do not need to seek the optimum de-anonymization scheme). Our objective is to derive the upper bound of the achievable anonymity and the lower bound of the achievable de-anonymity with statistical guarantee. Furthermore, in all the evaluations in this subsection, the default parameter settings are $\mu_1 = 0.7$, $\mu_0 = 0.9$, $\tau = 0.75$, and $\gamma = 0.02$.

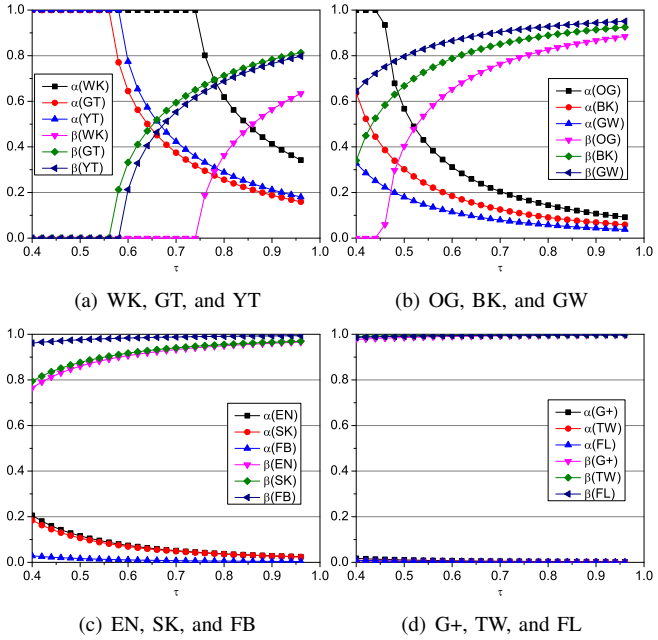
AUD vs. μ. First, we evaluate the anonymity and de-anonymity of the datasets in Table I with respect to the utility (characterized by μ_1 and μ_0) preserved by G^a. Due to the space limitations, we show the results of AUD vs. μ_1 in Fig.3 while omitting the results of AUD vs. μ_0. From Fig.3, we have three observations.

First, when μ_1 increases, the de-anonymity of each dataset increases while the anonymity of each dataset decreases. This is because μ_1 indicates the degree of G^a on preserving the existing relationships of G^r. A high μ_1 implies that G^a is more structurally similar to G^r, and thus is more structurally similar to G^u (for a given τ and γ). Therefore, more users in G^a are de-anonymizable leveraging the structural similarity between G^a and G^u.

Second, generally, the datasets with high d_{avg} (average degree) are more de-anonymizable (less anonymous) than that with low d_{avg}. This is because a higher d_{avg} implies richer local structural information is available in both G^a and G^u for de-anonymizing each user on average. Thus, a user is more likely to be correctly de-anonymized by structure-based de-anonymization attacks.

Finally, both the anonymity and the de-anonymity of graph data may exhibit the percolation phenomena[4], i.e., when μ_1 is below some threshold value, a graph achieves almost perfect anonymity; while when μ_1 is above some threshold value, an obvious loss of the anonymity happens. This implies that the actual anonymization/de-anonymization performance is sensitive to the utility carried by G^a and the structural similarity between G^a and G^u. Some increase on the similarity of G^a and G^u can induce a significant loss (resp., improvement) of the graph anonymity (resp., de-anonymity).

[4]It has been observed in [15], [22], the number of de-anonymizable users in seed-based two-phase de-anonymization attacks may exhibit the percolation phenomena with respect to the number of available seeds, i.e., when the number of seeds is below some threshold value, only a few users can be correctly de-anonymized; while when the number of seeds is above some threshold value, a significant portion of users are de-anonymizable.



(a) WK, GT, and YT (b) OG, BK, and GW

(c) EN, SK, and FB (d) G+, TW, and FL

Fig. 4. AUD vs. τ.

AUD vs. τ and γ. When τ increases, the anonymity and de-anonymity of each dataset are shown in Fig.4, from which we have two observations.

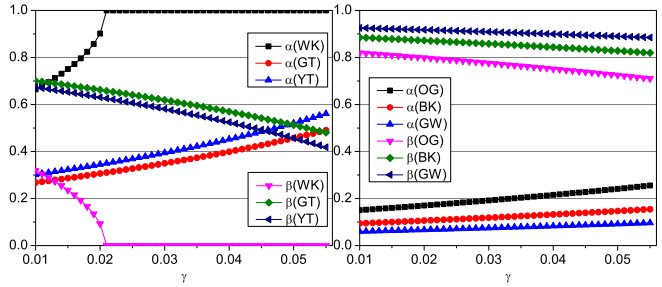
First, with the increase of τ, the anonymity (resp., de-anonymity) of each dataset decreases (resp., increases). This is because τ indicates how similar G^u and G^r are with respect to the existing relationships in G^r. Thus, a high τ implies G^u is more structurally similar to G^r and thus to G^a (when μ_1, μ_0, and γ are given), followed by G^a is more de-anonymizable by structure-based de-anonymization attacks leveraging G^u.

Second, again, the datasets with high d_{avg} are more de-anonymizable than those with low d_{avg}. The reason is also the same as we analyzed in Fig.3. A higher d_{avg} implies richer local structural information is available, which further enables more effective structure-based de-anonymization. In addition, similar to that in Fig.3, the anonymity and de-anonymity of a dataset may exhibit the percolation phenomena with respect to τ, e.g., Wiki, Gnutella, YouTube, and Oregon.

We show AUD vs γ in Fig.5, from which we have three observations.

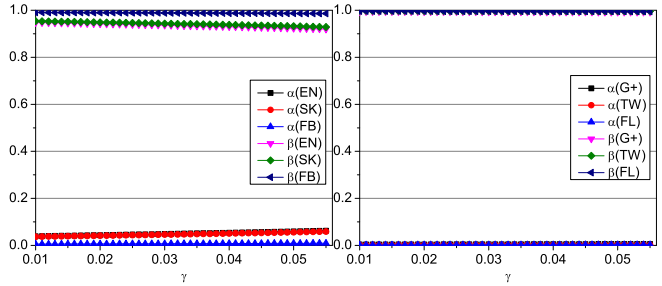
First, when γ increases, the anonymity of each dataset increases while the de-anonymity of each dataset decreases. This is because γ indicates the difference of G^r and G^u on users' non-existing relationships. Therefore, a large γ implies more structural difference between G^r and G^u with respect to the non-existing relationships, followed by more structural difference between G^a and G^u. Hence, less structural information can be leveraged to conduct successful de-anonymization and the anonymity of G^a is increased.

Second, generally, the anonymity and de-anonymity of datasets with low d_{avg} are more sensitive to the change of γ than that of datasets with high d_{avg}. This is because for



(a) WK, GT, and YT

(b) OG, BK, and GW



(c) EN, SK, and FB

(d) G+, TW, and FL

Fig. 5. AUD vs. γ.

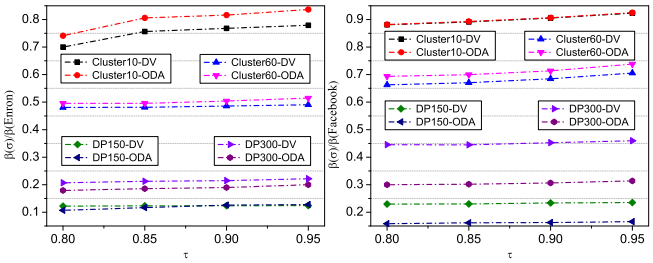
graphs with lower d_{avg}, the available structural information for de-anonyming each user is relatively less, and thus the structural/edge difference between G^a and G^r has more impacts on the achievable anonymity and de-anonymity.

Finally, similar as the results in Fig.3 and Fig.4, the anonymity and de-anonymity of a dataset may exhibit the percolation phenomena (e.g., Wiki).

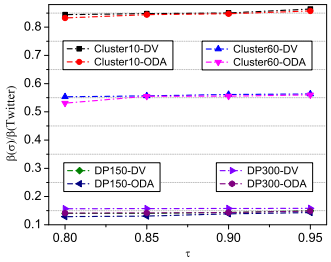
VIII. AUD-based Evaluation of State-of-the-Art Techniques

Methodology In this section, we conduct an AUD-based evaluation of the performance of state-of-the-art graph anonymization and de-anonymization techniques. The evaluation methodology is as follows: (i) given some graph datasets, anonymizing these datasets using state-of-the-art anonymization techniques; (ii) employing state-of-the-art de-anonymization attacks to de-anonymize the anonymized data and studying the data's practical de-anonymity; (iii) employing our AUD quantification technique to quantify the theoretical de-anonymity (anonymity) of the anonymized data; and (iv) finally, analyzing the practical and theoretical de-anonymity of the anonymized data.

Evaluation Setting. Here, we use three example datasets Enron, Facebook and Twitter as shown in Table I for this group of evaluation. The employed anonymization techniques are the latest cluster-based anonymization technique [8], denoted by Cluster, and the latest differential privacy-based anonymization technique [9], [10], denoted by DP. As we summarized in Section II, Cluster is a technique to make the users within a cluster have same local structures, and DP is a technique to make the dK-series of the anonymized graph meet a DP requirement. The employed de-anonymization attacks are the Distance-Vector (DV) based scheme proposed in [17] and the



(a) Enron (b) Facebook



(c) Twitter

Fig. 6. AUD-based Evaluation of state-of-the-art anonymization and de-anonymization techniques.

Optimization-based De-Anonymization (ODA) scheme proposed in [21]. As summarized in Section II, DV is a powerful seed-based de-anonymization attack while ODA is the latest seed-free de-anonymization attack.

When anonymizing the datasets, the key anonymization parameter for Cluster is the cluster size ζ [8] and for DP is the differential privacy parameter ξ [9], [10]. Basically, a larger ζ indicates a higher anonymization level for Cluster while a smaller ξ indicates a higher anonymization level for DP. In our evaluation, we consider the scenarios of $\zeta = 10$ and $\zeta = 60$ for Cluster and $\xi = 150$ and $\xi = 300$ for DP (which are similar to the settings in [8], [9]), respectively. For DV, since it requires seeds to bootstrap the de-anonymization, we randomly select 50 seed mappings from G^a to G^u in each evaluation. During the de-anonymization evaluation, the auxiliary datasets are obtained using a random edge adding/deleting process according to the specified τ and γ. In all the evaluations, we set $\gamma = \frac{\overline{\tau}\cdot|E^r|}{|E^r|}$. Furthermore, the required parameters μ_1 and μ_0 for AUD quantification can be obtained according to their definitions given G^r and G^a. For each group of evaluation, it will be repeated 50 times and the result is the average of the 50 runs.

Results. Let σ be a de-anonymization attack (e.g., DV and ODA) and n_c be the number of users that are successfully de-anonymized under σ. Then, the practical de-anonymity of an anonymized graph under σ is defined as $\beta(\sigma) = \frac{n_c}{n}$. Furthermore, to be consistent with previous notations, we use $\beta(\cdot)$ (e.g., $\beta(\text{Enron})$) to denote the AUD-based de-anonymity of a dataset, i.e., the theoretical de-anonymity. Then, we show $\frac{\beta(\sigma)}{\beta(\text{Enron})}$, $\frac{\beta(\sigma)}{\beta(\text{Facebook})}$, and $\frac{\beta(\sigma)}{\beta(\text{Twitter})}$ under different anonymization/de-anonymization scenarios in Fig.6, where "Cluster" and "DP" represent the anonymization algorithms, 10, 60, 150, and 300 represent the anonymization parameters, and DV and ODA represent the de-anonymization attack-

s, respectively. For instance, Cluster10-DV means that the anonymization algorithm applied is Cluster, the anonymization parameter used is 10, and the employed de-anonymization attack is DV. From Fig.6, we have three observations.

First, when τ increases, $\frac{\beta(\sigma)}{\beta(\cdot)}$ also has some increase, which implies that both DV and ODA can de-anonymize more users regardless of whether the dataset is anonymized by Cluster or DP. This because a large τ implies G^a and G^u are more structurally similar and thus G^u is more structurally similar to G^a. Therefore, more anonymized users can be successfully de-anonymized based on the structural information.

Second, for the scenarios of using Cluster as the anonymization algorithm, Cluster60 achieves better anonymity than Cluster10. This is because more users are made structurally similar under Cluster60 than that of Cluster10. However, intuitively, Cluster60 also sacrifices more data utility. Similarly, DP150 achieves better anonymity than DP300 at the cost of sacrificing more data utility. Overall, the datasets anonymized by DP achieves a better anonymity than that of Cluster. This is because DP changes more structural information of G^a than that of Cluster, i.e., the datasets anonymized by Cluster achieves a better utility than DP.

Finally and interestingly, there is still significant room for state-of-the-art de-anonymization techniques to be improved. From Fig.6, we have $\frac{\beta(\sigma)}{\beta(\cdot)} < 0.95$ in all the scenarios. Specifically, in the scenarios where DP is used, we have $\frac{\beta(\sigma)}{\beta(\text{Enron})} < 0.25$, $\frac{\beta(\sigma)}{\beta(\text{Facebook})} < 0.5$, and $\frac{\beta(\sigma)}{\beta(\text{Twitter})} < 0.16$ for both DV and ODA. Note that, according to our quantification, $\beta(\cdot)$ is only the lower bound of the de-anonymity of an anonymized graph. Therefore, the practical de-anonymity achieved by state-of-the-art de-anonymization attacks is much lower than the achievable theoretical de-anonymity, i.e., theoretically, significant room exists to improve existing de-anonymization attacks.

IX. Conclusion

In this paper, we conduct the first AUD correlation quantification for anonymized graph data and obtain close-forms under both the ER model and a general data model. Then, to facilitate the application of our quantification technique and to address the limitations of SecGraph, we integrate our AUD quantification to SecGraph as a new evaluation module and implement SecGraph+. Third, based on our quantification, we conduct a large scale evaluation on the anonymity, utility, and de-anonymity of real world graph data leveraging 12 datasets that are generated from various computer systems and services. Finally, we evaluate the performance of state-of-the-art anonymization and de-anonymization techniques in terms of our AUD quantification. We find that there is still significant space to improve existing de-anonymization attacks.

Acknowledgment

This work was partly supported by NSFC under No. 61772466, the Provincial Key Research and Development Program of Zhejiang, China under No. 2017C01055, the Fundamental Research Funds for the Central Universities, the Alibaba-Zhejiang University Joint Institute of Frontier Technologies, the CCF-NSFOCUS Research Fund under No. CCF-NSFOCUS2017011, and the CCF-Venustech Research Fund under No. CCF-VenustechRP2017009.

References

[1] S. Ji, P. Mittal, and R. Beyah, *Graph Data Anonymization, De-Anonymization Attacks, and De-Anonymizability Quantification: A Survey*, IEEE Comm. Surv. & Tutor., Vol. 19, No. 2, pp. 1305 - 1326, 2017.
[2] X. Ying and X. Wu, *Randomizing Social Networks: a Spectrum Preserving Approach*, SDM 2008.
[3] B. Zhou and J. Pei, *Preserving Privacy in Social Networks Against Neighborhood Attacks*, ICDE 2008.
[4] K. Liu and E. Terzi, *Towards Identity Anonymization on Graphs*, SIGMOD 2008.
[5] L. Zou, L. Chen, and M. T. Özsu, *K-Automorphism: A General Framework for Privacy Preserving Network Publication*, VLDB 2009.
[6] J. Cheng, A. Fu, and J. Liu, *K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks*, SIGMOD 2010.
[7] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, *Resisting Structural Re-identification in Anonymized Social Networks*, VLDB 2008.
[8] B. Thompson and D. Yao, *The Union-Split Algorithm and Cluster-based Anonymization of Social Networks*, ASIACCS 2009.
[9] A. Sala, X. Zhao, C. Wilson, H. Zheng, and B. Zhao, *Sharing Graphs using Differentially Private Graph Models*, IMC 2011.
[10] D. Proserpio, S. Goldberg, and F. McSherry, *Calibrating Data to Sensitivity in Private Data Analysis*, VLDB 2014.
[11] Y. Wang and X. Wu, *Preserving Differential Privacy in Degree-Correlation based Graph Generation*, TDP 2013.
[12] Q. Xiao, R. Chen, and K. Tan, *Differentially Private Network Data Release via Structural Inference*, KDD 2014.
[13] P. Mittal, C. Papamanthou, and D. Song, *Preserving Link Privacy in Social Network based Systems*, NDSS 2013.
[14] L. Backstrom, C. Dwork, and J. Kleinberg, *Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, WWW 2007.
[15] A. Narayanan and V. Shmatikov, *De-anonymizing Social Networks*, S&P 2009.
[16] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, *Community-enhanced De-anonymization of Online Social Networks*, CCS 2014.
[17] M. Srivatsa and M. Hicks, *Deanonymizing Mobility Traces: Using Social Networks as a Side-Channel*, CCS 2012.
[18] S. Ji, W. Li, M. Srivatsa, J. He, and R. Beyah, *Structure based Data De-anonymization of Social Networks and Mobility Traces*, ISC 2014.
[19] P. Pedarsani and M. Grossglauser, *On the Privacy of Anonymized Networks*, KDD 2011.
[20] N. Korula and S. Lattanzi, *An Efficient Reconciliation Algorithm for Social Networks*, VLDB 2014.
[21] S. Ji, W. Li, M. Srivatsa, and R. Beyah, *Structural Data De-anonymization: Quantification, Practice, and Implications*, CCS 2014.
[22] 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.
[23] The Google+ Dataset, http://www.cs.berkeley.edu/∼stevgong/dataset.html.
[24] Stanford Large Network Dataset Collection, *http://snap.stanford.edu/data/index.html*.
[25] ASU Social Computing Data Repository, *http://socialcomputing.asu.edu/pages/datasets*.
[26] http://idash.ucsd.edu/.
[27] G. J. Wills, NicheWorksInteractive Visualization of Very Large Graphs, Journal of Computational and Graphical Statistics, 1999.
[28] B. Hayes, Connecting the dots: Can the Tools of Grepha Theory and Social-network Studies Unravel the Next Big Plot? American Scientist, 2006.
[29] http://techcrunch.com/2014/10/29/twitter-partners-with-ibm-to-bring-social-data-to-the-enterprise/
[30] http://www.slideshare.net/jlcaut/ebola-hemoragic-fever-propagation-in-a-modern-city-using-sir-model
[31] 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.
[32] SecGraph, http://www.secgraph.gatech.edu/.