On the Relative De-anonymizability of Graph Data: Quantification and Evaluation

Shouling Ji†, Weiqing Li†, Shukun Yang†, Prateek Mittal‡, and Raheem Beyah†
† School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0765, USA
‡ Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA
Email: {sji, wli64, syang87}@gatech.edu, pmittal@princeton.edu, rbeyah@ece.gatech.edu

***Abstract*—In this paper, we propose a *structural importance-aware* approach to quantify the vulnerability/de-anonymizability of graph data to structure-based De-Anonymization (DA) attacks [1][2][3][4]. Specifically, we quantify both the *seed-based* and the *seed-free Relative De-anonymizability* (RD) of graph data for both perfect DA (successfully de-anonymizing all the target users) and partial DA (where some DA error is tolerated) under a general data model. In our relative quantification, instead of treating all the users in graph data as structurally equivalent, we adaptively quantify their RD in terms of their structural importance. Leveraging 15 real world graph datasets, we validate the accuracy of our relative quantifications and compare them with state-of-the-art seed-based and seed-free quantification techniques. The results demonstrate that our structural importance-aware relative quantifications are more sound and precise when measuring graph data's real vulnerability/de-anonymizability.**

I. Introduction

Graph data (e.g., social network data, contact data, healthcare data) are critical for academic research, government applications, commercial collaborations, healthcare applications, and other data mining tasks [1][5][6][7]. Therefore, graph data are frequently shared with researchers, government agencies, commercial parters, and other individuals/organizations [1][5][6][7]. One of the most notable characteristics of graph data is that the data items are structurally correlated with each other in addition to the semantic information they carried [1][5][6][7]. For instance, a user of a social network is correlated with other users in the network in addition to the profile information associated with him/her. On one hand, the correlations carried by graph data make the data useful for comprehensive analysis and meaningful applications (e.g., commercial and healthcare data mining and analysis tasks [1]). On the other hand, these correlations allow graph data to suffer security and privacy threats since adversaries can leverage them to infer private information about the users/systems who generated the graph data. As shown by recent research, anonymized graph data can be successfully de-anonymized in large-scale by structure-based De-Anonymization (DA) attacks [1][5][6][7].

To understand "the underlying reason for the success of existing structure-based DA attacks", recently, the concept of graph data de-anonymizability quantification has garnered significant research [5][8][9][10][11], where researchers study the following: based on graph data's structural information, why graph data can be de-anonymized, what are the DA conditions, and how many users are de-anonymizable, i.e., graph data de-anonymizability quantification can quantitatively examine how vulnerable/de-anonymizable any (anonymized) graph dataset is given its structure. Therefore, graph data de-anonymizability quantification techniques can be employed to examine the theoretically achievable vulnerability of both raw and anonymized graph data, and can thus evaluate the effectiveness of an anonymization scheme. Furthermore, the quantification results can serve as auxiliary information that is useful for future anonymization technique design and DA attack evaluation.

However, existing de-anonymizability quantifications [5][8][9][10][11] are limited due to some of the following reasons. First, most existing quantification techniques are either fully based on seed information, e.g., [8][10][11], or do not consider seed information at all, e.g., [5][9], which is an incomplete approach to graph data de-anonymizability analysis. Second, most existing quantification techniques are based on an impractical data model (e.g., the Erdös-Rényi (ER) model) and/or make impractical assumptions (e.g., dense seeds are available). Finally, all the existing techniques do not take into account the structural/topological importance of users in their quantifications. In practice, different users may have very different structural importance, e.g., the users with the maximum and minimum degrees are structurally different [2][12]. Therefore, existing quantification results are incomplete with regards to quantifying graph data's actual/precise structure-based de-anonymizability, i.e., they are inaccurate in quantifying graph data's actual vulnerability to structure-based DA attacks (we summarize existing de-anonymizability quantification techniques and discuss their limitations in detail in Sections II).

Contributions. To address the limitations of existing techniques, we study the *structural importance-aware Relative De-anonymizability (RD)* of graph data in this paper. Instead of treating all the users as equivalent, we quantify the de-anonymizability of anonymized graph data adaptively and accurately by taking into account users' structural differences. Specifically, our contributions are summarized as follows.

1) We introduce the concept of *structural importance-aware RD* of graph data. Specifically, we formally define θ-*RD* and (θ, ϵ)-*RD* of graph data, where θ indicates the target DA users according to their structural importance and ϵ characterizes the tolerated *DA error* by a DA scheme.

978-1-4673-9953-1/16/$31.00 ©2016 IEEE

2) Under a general data model, we quantify the *seed-based RD* of graph data for both perfect DA (de-anonymizing all the target users) and partial DA (where some DA error is tolerated). Our seed-based quantification provides the most accurate theoretical foundation for the success of existing seed-based DA attacks (e.g.,[1][2][3]).
3) Under a general data model, we also quantify the *seed-free RD* of graph data for both perfect DA and partial DA. Our seed-free quantification provides the most accurate theoretical foundation for the success of existing seed-free DA attacks (e.g., [4][5]).
4) Leveraging 15 real world graph datasets, we conduct a large-scale evaluation of our RD quantification techniques. We also evaluate the performance our quantifications against that of state-of-the-art quantification techniques. The evaluation results demonstrate that our structural importance-aware RD quantification techniques are more sound and accurate when measuring graph data's real vulnerability/de-anonymizability.

The reminder of this paper is as follows. In Section II, we summarize the related work. The system model and preliminaries are given in Section III. We quantify seed-based and seed-free RD of graph data in Sections IV and V, respectively. The large-scale evaluation is conducted in Section VI. We conclude this paper in Section VII.

II. Related Work

A. Graph Data DA

In [3], Backstrom et al. proposed both active and passive DA attacks to graph data based on subgraph pattern matching. Later, Narayanan and Shmatikov proposed the first scalable and robust two-phase DA attack in [1], where the first phase is used for seed identification and the second phase is for DA propagation. In addition, Narayanan et al. studied how to perform link DA in [13] leveraging node/user DA. In [14], Nilizadeh et al. extended the attack in [1] and presented a community-enhanced DA attack. In [6], Srivatsa and Hicks presented three two-phase attacks similar to [1] to de-anonymize contact graphs (constructed from mobility traces). In [2], Ji et al. proposed another seed-based two-phase DA framework. There are also seed-free DA attacks, e.g., in [5], a seed-free optimization-based DA attack is designed. In [7], Ji et al. developed SecGraph, a uniform and open-source evaluation system for graph anonymization and de-anonymization.

B. De-anonymizability Quantification

1) Seed-based Quantification: In [10], Yartseva and Grossglauser quantified the de-anonymizability of graph data by analyzing a percolation-based graph matching algorithm under the Erdös-Rényi (ER) random graph model $G(n,p)$ (a random graph consists of n nodes/users, and an edge exists between any pair of nodes with probability p). However, it is seldom to see that any real world graph data following the ER model (if not possible) [5][12], the quantification under the ER model is only mathematically meaningful but not practical. Nevertheless, it can shed light on more practical quantification. Another limitation of [10] is that it leverages seed-associated structural information for de-anonymizability quantification. In fact, as shown in [9][5], graph data is de-anonymizable based solely on data's structural information, i.e., without seeds.

Following the same direction, Korula and Lattanzi conducted another seed-based de-anonymizability quantification of graph data under both the ER model and the Preferential Attachment (PA) model [11]. However, the quantification in [11] is valid under a strong assumption of existing dense seeds ($\Theta(\iota \cdot n)$ available seeds, $\iota \in (0,1]$ is a constant), which is not true for real world DA attacks. Recently, Ji et al. also quantified the seed-based de-anonymizability of social networks [8] under the ER model and a statistical model.

2) Seed-free Quantification: In [9], Pedarsani and Grossglauser quantified the de-anonymizability of graph data under the ER model. Again, the quantification is under the mathematical ER model, which cannot be applied to real world graph data [5][12]. Ji et al. improved the quantification in [9]. Similar to [9], the authors did not employ seeds neither.

III. System Model, Definitions, and Preliminaries

In this section, we present the system model for RD quantification. To provide consistency, we employ the same data model, assumptions, and notations with existing quantification work [5][9][10][11].

A. Data Model

Since we focus on structure-based de-anonymizability quantification, we model the anonymized data as a graph denoted by $G^a = (V^a, E^a)$, where $V^a = \{i | i \text{ is an anonymized user}\}$ is the user set and $E^a = \{e^a_{i,j} | i,j \in V^a\}$ is the edge/link set [5][8][9][10][11]. To de-anonymize G^a, we assume there is an auxiliary graph $G^u = (V^u, E^u)$ available, where $V^u = \{i | i \text{ is a known user}\}$ and $E^u = \{e^u_{i,j} | i,j \in V^u\}$ are the user and edge/link sets, respectively [5][8][9][10][11]. As indicated in [1][5][6][8][13], G^u is widely available and can be obtained through multiple means, e.g., online crawling, data aggregation, regular data publishing by companies and government agencies, advertising, and third-party applications.

For facilitating our quantification, we restate the two assumptions in [5][8][9][10][11]. First, we assume $V^a = V^u$ although we do not know the exact mappings between the users in V^a and V^u. Note that, this assumption does not limit existing works [5][8][9][10][11] nor our quantification. In the case that $V^a \neq V^u$, we can simply make $V^a = V^u$ by adding users in $V^u \setminus V^a$ (resp., $V^a \setminus V^u$) to V^a (resp. V^u) as isolated nodes (with degree of 0). Second, following the first assumption, we assume that G^a and G^u are two sampled graphs of an underlaying *conceptual* graph $G = (V, E)$, where $V = V^a = V^u = \{i | i \text{ is a physical user}\}$ and $E = \{e_{i,j} | i,j \in V\}$ characterizes the possible real world physical relationships (edges/links) among the users in V. For instance, given a group of people, their Facebook network mainly carries their friendship relation while their LinkedIn network mainly carries their career connections. Mathematically, this assumption can be formalized as $\forall e_{i,j} \in E$,

$\Pr(e_{i,j} \in E^a) = s_a$ and $\Pr(e_{i,j} \in E^u) = s_u$, where s_a and s_u are the *sampling probabilities* of G^a and G^u, respectively. Note that, the second assumption is only for the mathematical purpose of conveniently quantifying the structural similarity between G^a and G^u and does not limit the generality of our quantification[1]. Even without this assumption, the derivations in [8][5][9][10][11] and this paper are still valid, however, they will be much more complicated (in that scenario, we need to define more functions to characterize and calculate the structural similarity between G^a and G^u).

Now, we define some useful notations. We assume $|V| = n$ and $|E| = m$. $\forall i \in V$, we define its *neighborhood* and *degree* as $N_i = \{j | e_{i,j} \in E\}$ and $d_i = |N_i|$, respectively. Without loss of generality, we assume $d_i \leq d_j$ for $i < j$, which implies $d_1 \leq d_2 \leq \cdots \leq d_n$. $\forall U \subseteq V$, we define $n_U = |U|$, $E_U = \{e_{i,j} \in E | i, j \in U\}$ be the set of edges among the users in U, and $m_U = |E_U|$. Furthermore, $\forall U \subseteq V$, we use δ_1^U and δ_2^U to denote its *smallest* and *second smallest* degrees of the users in U, respectively, and use Δ_1^U and Δ_2^U to denote its *largest* and *second largest* degrees of the users in U, respectively. For instance, when $U = V$, $\delta_1^U = d_1$, $\delta_2^U = d_2$, $\Delta_1^U = d_n$, and $\Delta_2^U = d_{n-1}$. Given $U, W \subseteq V$ and $U \cap W = \emptyset$, we define the set of cross edges between U and W as $E_{U,W} = \{e_{i,j} \in E | i \in U, j \in W\}$, and $m_{U,W} = |E_{U,W}|$.

To make our quantification general, as done in [5], we assume G follows the *configuration model* [12], under which G can have an *arbitrary* degree sequence that follows any distribution. For $\forall i, j \in V$, let $p_{i,j}$ be the probability of an edge existing between i and j. Then, following the key property of the configuration model, we have $p_{i,j} = \frac{d_i d_j}{2m-1} \underset{\text{as } m \to \infty}{\simeq} \frac{d_i d_j}{2m}$. For $\forall U \subseteq V$, we define $l_U = \min\{p_{i,j} | i, j \in U\}$ and $h_U = \max\{p_{i,j} | i, j \in U\}$ be the *minimum* and *maximum existing probabilities* of the edges in E_U, respectively. Furthermore, given $i \in V$, $U \subseteq V$, and $i \notin U$, we define $l_{i,U} = \min\{p_{i,j} | j \in U\}$ and $h_{i,U} = \max\{p_{i,j} | j \in U\}$ be the *minimum* and *maximum existing probabilities* of the edges in $E_{\{i\},U}$.

In the following discussion of this paper, to minimize confusion, we use i, $e_{i,j}$, d_i, N_i, and other notations interchangeably in the contexts of G, G^a, and G^u. For some specific scenarios, we use superscripts '*a*' and '*u*' to distinguish between the contexts of G^a and G^u, respectively.

B. Relative De-anonymizability

Now, we formally define RD. First, we formally define a DA attack. Given G^a and G^u, a *DA attack* is defined as a mapping [5], [8], [9]: $\sigma : V^a \to V^u$ as shown in Fig.1 (which indicates a mapping form G^a to G^u), i.e., $\sigma = \{(i, \sigma(i)) | i \in V^a\}$.

[1]In practice, G^a may be obtained by anonymizing the raw data using any graph anonymization technique. Meanwhile, G^u may be any specific auxiliary graph of the anonymized users. However, it is also reasonable to consider G^a and G^u are correlated and structurally similar [2][5][9][15]. For instance, if two users are connected on Facebook, they are also more likely to be connected on LinkedIn compared to two arbitrary users without any connection on Facebook. In consideration of this, we introduce and use the conceptual graph G to simplify the characterization and formalization of the structural similarity between G^a and G^u.

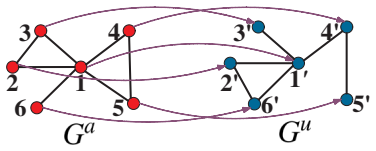


Fig. 1. A DA attack.

Under σ, $\forall i \in V^a$, $\sigma(i) \in V^u \cup \{\perp\}$, where $\perp$ is a special *not existing indicator*. Furthermore, $\forall e_{i,j}^a \in E^a$, let $\sigma(e_{i,j}^a) = e_{\sigma(i),\sigma(j)}$ and $\forall U \subseteq E^a$, $\sigma(U) = \{\sigma(e_{i,j}^a) | e_{i,j}^a \in U\}$. Since we assume $V^a = V^u$, we define $i \in V^a$ is successfully de-anonymized if $\sigma(i) = i$. Then, given a mapping σ, we say it is a *k-error DA*, denoted by σ_k, if there are k users that are incorrectly de-anonymized, i.e., exists k mappings $(i, \sigma(i))$ in σ such that $\sigma(i) \neq i$. Specifically, when $k = 0$, σ_0 is called the *perfect DA*.

As we discussed before, in a given G, different users have different structural importance for structural DA attacks. Therefore, instead of treating all the users the same as done in previous quantifications [5][8][9][10][11], we study this problem under a more general context, where we quantify the de-anonymizability of users adaptively according to their relative structural importance. According to the empirical results in [2], a user with a higher degree implies it has higher structural importance, e.g., closeness centrality, betweenness centrality [12]. Therefore, we measure users' structural importance according to their degrees. Then, we define G's *θ-RD* as follows.

Definition III.1. ***θ-RD and θ-relative DA***. *Given G, G^a, G^u and $\theta \in [0, 1]$, let $V_\theta = \{n, n-1, \cdots, (1-\theta)n + 1\}$ be the users in G with top θn degrees (without loss of generality, we assume θn is an integer value). Then, G^a (or equivalently, G) is* θ-relatively de-anonymizable *if all the users in V_θ are perfectly de-anonymizable based on their structural information carried by G^a and G^u. Furthermore, given σ, if all the users in V_θ are perfectly de-anonymized, σ is a* θ-relative DA.

From the above definition, we can see that when $\theta = 0$, θ-RD turns to be the de-anonymizability definition considered in [5][8]-[11]. Therefore, the quantification problem studied in [5][8]-[11] can be considered as a special case of θ-relative DA (we formally state this later). To make our quantification more applicable, we also consider the scenario that σ tolerates some *DA error*, indicated by $\epsilon \in [0, \theta]$, as follows.

Definition III.2. ***(θ, ϵ)-RD and (θ, ϵ)-relative DA***. *Given G, G^a, G^u, θ, and $\epsilon \in [0, \theta]$, G^a (or G) is* (θ, ϵ)-relative de-anonymizable *if at least $(\theta - \epsilon)n$ users in V_θ can be perfectly de-anonymized based on their structural information. Furthermore, σ is a* (θ, ϵ)-relative DA *if at least $(\theta - \epsilon)n$ users in V_θ are perfectly de-anonymized under σ.*

To measure the performance of a de-anonymization attack σ, we use the *edge difference* between G^a and G^u under σ [5], [8], [9], which is defined as $\Psi_\sigma = |\sigma(E^a) \setminus E^u| + |\sigma^{-1}(E^u) \setminus$

$E^a|$, i.e., Ψ_σ counts the number of edges that appeared in one graph (G^a/G^u) while not appearing in another graph (G^u/G^a). Furthermore, $\forall U \subseteq E$, the edge difference between G^a and G^u associated with edges in U under σ is defined as $\Psi_{\sigma:U} = |\sigma(U^a) \setminus U^u| + |\sigma^{-1}(U^u) \setminus U^a|$.

IV. Seed-based RD Quantification

In this section, we study the scenario that an adversary has some *seed knowledge*. We denote the known *seed set* as $S = \{i | i \in V, i \text{ is a seed}\}$ and let $\Lambda = |S|$. For simplicity, we first assume that G^a and G^u are sampled versions of G with the same sampling probability s, i.e., $s_a = s_u = s$. Then, we extend our quantification to the case that $s_a \neq s_u$.

A. θ-RD

For $\forall i \in V \setminus S$, let $l_i = l_{i,S}$, $h_i = h_{i,S}$. Define $f_s^s = \min_{i \in V_\theta \setminus S, j \in V \setminus (S \cup \{i\})} \frac{s(l_i(1-h_j s)+l_j(1-h_i s)-2h_i(1-s))^2}{8(l_i(1-h_j s)+l_j(1-h_i s)+2h_i(1-s))}$. We quantify the condition to de-anonymize a user given S as follows.

Theorem 1. *For $\forall i \in V \setminus S$, if $s > \max_{j \in V \setminus (S \cup \{i\})} \frac{2h_i - l_i - l_j}{2h_i - l_i h_j - l_j h_i}$ and $\Lambda = \Omega(\frac{2 \ln n + 1}{f_s^s})$, then it is* asymptotically almost surely *(*a.a.s.*)*[2] *that i is perfectly de-anonymizable given S.*

Proof Sketch: $\forall v \in S$, $\Pr(e_{i,v} \in E) = p_{i,v}$ under the configuration model. Furthermore, for mapping (i,i), $e_{i,v}$ causes an edge difference if it is sampled into one graph (G^a/G^u) while not into the other (G^u/G^a). Therefore, we have $\Pr(e_{i,v}$ induces one edge difference to mapping $(i,i)) = 2p_{i,v}s(1-s)$. Similarly, if i is incorrectly de-anonymized to some $j \neq i$, then we have $\Pr(e_{i,v}/e_{j,v}$ induces one edge difference to mapping $(i,j)) = p_{i,v}s(1-p_{j,v}s) + p_{j,v}s(1-p_{i,v}s)$. Let $E_{i,S} = \{e_{i,v} | v \in S\}$, X be a random variable counting the edge differences for mapping (i,i) caused by edges in $E_{i,S}$, and Y be a random variable counting the edge differences for mapping $(i, j \neq i)$ caused by edges in $E_{i,S}$ and $E_{j,S}$. Then, we have $X \sim \sum_{v \in S} \mathbf{B}(1, 2p_{i,v}s(1-s)) \leq \mathbf{B}(\Lambda, 2h_i s(1-s))$ and $Y \sim \sum_{v \in S} \mathbf{B}(1, p_{i,v}s(1-p_{j,v}s) + p_{j,v}s(1-p_{i,v}s)) \geq \mathbf{B}(\Lambda, l_i s(1-h_j s) + l_j s(1-h_i s))$, where $\mathbf{B}(\cdot,\cdot)$ is a *binomial distribution*. Let λ_X and λ_Y be the mean values of X and Y, respectively. Since $s > \max_{v \in S, j \in V \setminus S, j \neq i} \frac{p_{i,v} - p_{j,v}}{2p_{i,v}(1-p_{j,v})}$, we have $\lambda_Y > \lambda_X$. Then, applying the Pedarsani-Grossglauser Lemma [9], we have $\Pr(X \geq Y) \leq 2\exp(-\frac{(\lambda_Y - \lambda_X)^2}{8(\lambda_X + \lambda_Y)}) \leq 2\exp(-\Lambda f_s^s) \leq 2\exp(-2\ln n - 1) \leq \frac{1}{n^2}$. Since $\sum_{n \geq 1} \frac{\pi^2}{6} \leq \infty$, according to the *Borel-Cantelli Lemma*, $\Pr(X \geq Y) \underset{n \to \infty}{\to} 0 \Leftrightarrow \Pr(X < Y) \underset{n \to \infty}{\to} 1$, i.e., i is perfectly de-anonymizable based on the edge difference relative to S. □

Note that, the conclusion in Theorem 1 is valid not only for any user in $V_\theta \setminus S$ but also for any user in $V \setminus S$. Now, based on Theorem 1, we quantify a stronger conclusion, which indicates the condition to perfectly de-anonymize all the users in $V_\theta \setminus S$, i.e., the θ-*RD* of G^a. Let $\mathcal{E}^s$ be the event that *there is at least one incorrectly de-anonymized user in $V_\theta \setminus S$*. Then, we have the following theorem.

Theorem 2. *If $s > \max_{i \in V_\theta \setminus S, j \in V \setminus (S \cup \{i\})} \frac{2h_i - l_i - l_j}{2h_i - l_i h_j - l_j h_i}$ and $\Lambda = \Omega(\frac{2 \ln n + \ln 2\theta n}{f_s^s})$, it is a.a.s. that $\Pr(\mathcal{E}^s) \underset{n \to \infty}{\to} 0$, i.e., G^a is θ-relatively de-anonymizable.*

Proof Sketch: Based on the *Boole's inequality* and Theorem 1, we have $\Pr(\mathcal{E}^s) \leq \sum_{i \in V_\theta \setminus S} \Pr(i \text{ is incorrectly de-anonymized}) \leq \sum_{i \in V_\theta \setminus S} 2\exp(-\Lambda f_s^s) \leq 2\theta n \exp(-\Lambda f_s^s) \leq \frac{1}{n^2}$. Then, according to the *Borel-Cantelli Lemma*, we have $\Pr(\mathcal{E}^s) \underset{n \to \infty}{\to} 0$. □

B. (θ, ϵ)-RD

Now, we study the DA that tolerates some *DA error*, i.e., the (θ, ϵ)-*RD* of G^a. Let $V_{\theta-\epsilon} = \{n, n-1, \cdots, (1+\epsilon-\theta)n+1\}$. Then, we demonstrate the quantification in Theorem 3.

Theorem 3. *If $s > \max_{i \in V_{\theta-\epsilon} \setminus S, j \in V \setminus (S \cup \{i\})} \frac{2h_i - l_i - l_j}{2h_i - l_i h_j - l_j h_i}$ and $\Lambda = \Omega(\frac{2 \ln n + \ln 2(\theta-\epsilon)n}{f_s^s})$, it is a.a.s. that G^a is (θ, ϵ)-relatively de-anonymizable.*

Proof Sketch: To prove this theorem, it is sufficient to quantify the condition under which there exists a perfectly de-anonymizable subset of V_θ of size at least $(\theta - \epsilon)n$. Here, we consider $V_{\theta-\epsilon}$ since the users in this set have higher structural importance, i.e., they carry more structural information. Let $\mathcal{E}_\epsilon^s$ be the event that *there is at least one incorrectly de-anonymized user in $V_{\theta-\epsilon}$*. Thus, it is sufficient to prove that it is a.a.s. $\Pr(\mathcal{E}_\epsilon^s) \underset{n \to \infty}{\to} 0$. Then, based on similar arguments as in Theorem 2, this theorem can be proven. □

C. Extension: $s_a \neq s_u$

In the previous quantification, we assume $s_a = s_u = s$. In reality, it is possible that $s_a \neq s_u$. Now, we quantify the θ-*RD* and (θ, ϵ)-*RD* of G^a in this case. Let $f_{s_a,s_u}^s = \min_{i \in V_\theta \setminus S, j \in V \setminus (S \cup \{i\})} \frac{((2h_i - l_i h_j - l_j h_i)s_a s_u + (l_i - h_i)s_a + (l_j - h_i)s_u)^2}{8((l_i + h_i)s_a + (l_j + h_i)s_u - (2h_i + l_i h_j + l_j h_i)s_a s_u)}$. We have the following theorem. The proof is omitted due to the space limitations.

Theorem 4. *Suppose $\frac{s_a s_u}{s_a + s_u} > \max_{i \in V_\theta \setminus S, j \in V \setminus (S \cup \{i\})} \frac{h_i - \min\{l_i, l_j\}}{2h_i - l_i h_j - l_j h_i}$. Then, (i) if $\Lambda = \Omega(\frac{2 \ln n + 1}{f_{s_a,s_u}^s})$, it is a.a.s. that i is perfectly de-anonymizable given S; (ii) if $\Lambda = \Omega(\frac{2 \ln n + \ln 2\theta n}{f_{s_a,s_u}^s})$, it is a.a.s. that G^a is θ-relatively de-anonymizable; and (iii) if $\Lambda = \Omega(\frac{2 \ln n + \ln 2(\theta-\epsilon)n}{f_{s_a,s_u}^s})$, it is a.a.s. that G^a is (θ, ϵ)-relatively de-anonymizable.*

V. Seed-free RD Quantification

In the previous section, we studied seed-based RD quantification. In this subsection, we study *seed-free RD quantifica-*

[2] *Asymptotically almost surely* (*a.a.s.*) implies that as $n \to \infty$, with probability goes to 1 an event happens.

tion. Again, for simplicity, we first assume $s_a = s_u = s$ and then extend to the case that $s_a \neq s_u$.

A. θ-RD

Let $l = l_V$, $l_\theta = l_{V_\theta}$, $h_\theta = h_{V_\theta}$, $\kappa = (1-\theta)n$ and σ_κ be a *θ-relative DA*. Furthermore, let $k \in [1, \theta n]$ and $\sigma_{\kappa+k}$ be a DA such that k users are incorrectly de-anonymized. Define $f_s = \frac{s((l+l_\theta)(1-h_\theta s)-2h_\theta(1-s))^2}{8((l+l_\theta)(1-h_\theta s)+2h_\theta(1-s))}$. We have the following theorem.

Theorem 5. *If $s > \frac{2h_\theta - l_\theta - l}{h_\theta(2-l_\theta-l)}$ and $f_s = \Omega(\frac{2\ln n+1}{\theta k n - k^2/2 - k})$, then statistically, $\Pr(\Psi_{\sigma_\kappa} < \Psi_{\sigma_{\kappa+k}}) \underset{n\to\infty}{\to} 1$, i.e., it is a.a.s. that a θ-relative DA induces less edge difference than any DA which incorrectly de-anonymizes some users in V_θ.*

Proof Sketch: Let $\overline{E_{V_\theta}} = E \setminus E_{V_\theta}$. Then, according to the edge difference definition, we have $\Psi_\sigma = \Psi_{\sigma:E_{V_\theta}} + \Psi_{\sigma:\overline{E_{V_\theta}}}$ for any σ. Since we only focus on de-anonymizing the users in V_θ, statistically, to prove $\Pr(\Psi_{\sigma_\kappa} < \Psi_{\sigma_{\kappa+k}}) \underset{n\to\infty}{\to} 1$, it is sufficient to prove $\Pr(\Psi_{\sigma_\kappa:E_{V_\theta}} < \Psi_{\sigma_{\kappa+k}:E_{V_\theta}}) \underset{n\to\infty}{\to} 1$. This is because $\Psi_{\sigma_\kappa:\overline{E_{V_\theta}}} \underset{statistically, n\to\infty}{=} \Psi_{\sigma_{\kappa+k}:\overline{E_{V_\theta}}}$.

Let $X = \Psi_{\sigma_\kappa:E_{V_\theta}}$ and $Y = \Psi_{\sigma_{\kappa+k}:E_{V_\theta}}$ be two random variables. Furthermore, under $\sigma_{\kappa+k}$, let $C \subset V_\theta$ and $I \subseteq V_\theta$ be the sets of users that are correctly and incorrectly de-anonymized, respectively. Evidently, $C \cup I = V_\theta$, $C \cap I = \emptyset$, and $E_{V_\theta} = E_C \cup E_I \cup E_{C,I}$. For $e_{i,j} \in E_I$, it is a *transposition edge* if $(i,j),(j,i) \in \sigma_{\kappa+k}$. Let $E_t = \{e_{i,j} | e_{i,j}$ is a transposition edge under $\sigma_{\kappa+k}\}$ and $m_t = |E_t|$. Then, $m_t \leq k/2$. Let $m_\theta = |E_\theta| = \binom{n}{2}$, $m_C = |E_C| = \binom{\theta n}{\theta n - k}$, $m_I = |E_I| = \binom{k}{2}$, and $m_{C,I} = |E_{C,I}| = k(\theta n - k)$. Now, $\forall e_{i,j} \in E_{V_\theta}$, it contributes one to X if it is existed and sampled into exactly one of G^a and G^u. Thus, $X \sim \sum_{e_{i,j}\in E_{V_\theta}} \mathbf{B}(1, 2p_{i,j}s(1-s))$. Similarly, $\forall e_{i,j} \in E_C \cup E_t$, it contributes one to Y if it is existed and sampled into exactly one of G^a and G^u. $\forall e_{i,j} \in E_{C,I} \cup (E_I \setminus E_t)$, it contributes one to Y if $e_{i,j}$ (or, $\sigma(e_{i,j})$) is appeared in G^a (or, G^u) while $\sigma(e_{i,j})$ (or, $e_{i,j}$) is not appeared in G^u (or, G^a). Let $q_{i,j} = p_{\sigma_{\kappa+k}(i),\sigma_{\kappa+k}(j)}$, we have $Y \sim \sum_{e_{i,j}\in E_C \cup E_t} \mathbf{B}(1, 2p_{i,j}s(1-s)) + \sum_{e_{i,j}\in E_{C,I}\cup(E_I\setminus E_t)} \mathbf{B}(1, p_{i,j}s(1-q_{i,j}s) + q_{i,j}s(1-p_{i,j}s))$. Let $X' \sim \sum_{e_{i,j}\in E_{C,I}\cup(E_I\setminus E_t)} \mathbf{B}(1, 2p_{i,j}s(1-s)) \leq \mathbf{B}(m_{C,I} + m_I - m_t, 2h_\theta s(1-s))$ and $Y' \sim \sum_{e_{i,j}\in E_{C,I}\cup(E_I\setminus E_t)} \mathbf{B}(1, p_{i,j}s(1-q_{i,j}s) + q_{i,j}s(1-p_{i,j}s)) \geq \mathbf{B}(m_{C,I} + m_I - m_t, s(l+l_\theta)(1-h_\theta s))$. Thus, $\Pr(X < Y) = \Pr(X' < Y')$. Let $\lambda_{X'}$ and $\lambda_{Y'}$ be the mean values of X' and Y', respectively. Then, since $s > \frac{h_\theta - l}{h_\theta(1-l)}$, $\lambda_{X'} < \lambda_{Y'}$. Thus, based on the Pedarsani-Grossglauser Lemma [9], $\Pr(X' \geq Y') \leq 2\exp(-\frac{(\lambda_{Y'}-\lambda_{X'})^2}{8(\lambda_{X'}+\lambda_{Y'})}) \leq 2\exp(-(\theta k n - k^2/2 - k) f_s) \leq 2\exp(-2\ln n - 1) \leq \frac{1}{n^2}$. Thus, according to the *Borel-Cantelli Lemma*, $\Pr(X' \geq Y') \underset{n\to\infty}{\to} 0$, i.e., $\Pr(X' < Y') = \Pr(X < Y) \underset{statistically, n\to\infty}{=} \Pr(\Psi_{\sigma_\kappa} < \Psi_{\sigma_{\kappa+k}}) \underset{n\to\infty}{\to} 1$. □

In theorem 5, we quantified the condition such that $\Pr(\Psi_{\sigma_\kappa} < \Psi_{\sigma_{\kappa+k}})$ for any $\sigma_{\kappa+k}$. Now, we quantify the *θ-RD* of G^a. To achieve this, statistically, we need to show the *uniqueness* of σ_κ, i.e., $\nexists \sigma_{\kappa+k}$ such that $\Psi_{\sigma_{\kappa+k}} \leq \Psi_{\sigma_\kappa}$. We give the quantification in Theorem 6. The proof is omitted due to the space limitations.

Theorem 6. *If $s > \frac{2h_\theta - l_\theta - l}{h_\theta(2-l_\theta-l)}$ and $f_s = \Omega(\frac{2\ln n + (k+1)\ln\theta n + 1}{\theta k n - k^2/2 - k})$, then statistically, it is a.a.s. that $\nexists \sigma_{\kappa+k}$ such that $\Psi_{\sigma_{\kappa+k}} \leq \Psi_{\sigma_\kappa}$, i.e., it is a.a.s. that G^a is θ-relatively de-anonymizable.*

B. (θ, ϵ)-RD

Now, we study the RD of G^a when some error is tolerated. We show the quantification in Theorem 7, where the proof is omitted due to the space limitations.

Theorem 7. *If $s > \frac{2h_{\theta-\epsilon} - l_{\theta-\epsilon} - l}{h_{\theta-\epsilon}(2-l_{\theta-\epsilon}-l)}$ and $f_s = \Omega(\frac{2\ln n + (k+1)\ln(\theta-\epsilon)n + 1}{(\theta-\epsilon)kn - k^2/2 - k})$, it is a.a.s. that G^a is (θ,ϵ)-relatively de-anonymizable.*

C. Extension: $s_a \neq s_u$

In Theorems 5, 6, and 7, we consider the case that $s_a = s_u$. When $s_a \neq s_u$, let $f_{s_a,s_u} = \frac{((2-l-l_\theta)h_\theta s_a s_u + (l_\theta - h_\theta)s_a + (l-h_\theta)s_u)^2}{8((l_\theta+h_\theta)s_a + (l+h_\theta)s_u - (2+l+l_\theta)h_\theta s_a s_u)}$. We quantify the RD of G^a as follows. The proof is omitted due to the space limitations.

Theorem 8. *Statistically, (i) if $\frac{s_a s_u}{s_a+s_u} > \frac{h_\theta - l}{h_\theta(2-l_\theta-l)}$ and $f_{s_a,s_u} = \Omega(\frac{2\ln n+1}{\theta k n - k^2/2 - k})$, $\Pr(\Psi_{\sigma_\kappa} < \Psi_{\sigma_{\kappa+k}}) \underset{n\to\infty}{\to} 1$; (ii) if $\frac{s_a s_u}{s_a+s_u} > \frac{h_\theta - l}{h_\theta(2-l_\theta-l)}$ and $f_{s_a,s_u} = \Omega(\frac{2\ln n + (k+1)\ln\theta n + 1}{\theta k n - k^2/2 - k})$, it is a.a.s. that G^a is θ-relatively de-anonymizable; and (iii) if $\frac{s_a s_u}{s_a+s_u} > \frac{h_{\theta-\epsilon} - l}{h_{\theta-\epsilon}(2-l_{\theta-\epsilon}-l)}$ and $f_{s_a,s_u} = \Omega(\frac{2\ln n + (k+1)\ln(\theta-\epsilon)n + 1}{(\theta-\epsilon)kn - k^2/2 - k})$, it is a.a.s. that G^a is (θ,ϵ)-relatively de-anonymizable.*

TABLE I
DATA STATISTICS.

Name	Type	n	m	ρ	$\overline{d}$
Google+	SN	4,692,671	90,751,480	8.24E-6	38.7
LiveJournal	SN	4,847,571	68,993,773	3.70E-6	17.9
YouTube	SN	1,134,890	2,987,624	4.64E-6	5.3
Orkut	SN	3,072,441	117,185,083	2.48E-5	76.3
Pokec	SN	1,632,803	30,622,564	1.67E-5	27.3
Facebook	SN	63,731	817,090	4.02E-4	25.6
Flickr	SN	80,513	5,899,882	1.82E-3	146.6
Foursquare	SN	639,014	3,214,986	1.57E-5	10.1
Twitter	SN	456,631	14,855,875	1.20E-4	54.8
Gowalla	LSN	196,591	950,327	4.92E-5	9.7
AstroPh	Collab.	18,772	396,160	1.23E-3	22.0
Enron	Email	36,692	183,831	3.19E-4	10.7
EuAll	Email	265,214	420,045	1.35E-5	3.0
Skitter	AS	1,696,415	11,095,298	7.73E-6	13.1
Gnutella	P2P	26,518	65,369	1.86E-4	4.9

VI. LARGE-SCALE EVALUATION

A. Datasets

To validate our quantification techniques, we employ 15 real world graph datasets (which are obtained from Stanford SNAP [16]/ASU Data Repository [17], and have been used in the

latest quantification work [5], [8]) as shown in Table I, where n indicates the number of nodes (users), m is the number of edges (links/relationships), ρ is the *graph density*, and $\overline{d}$ is the average degree of each node. The 15 datasets are generated from various computer systems/services. We briefly introduce the datasets as follows.

• **Social Network Graph (SN).** *Google+* is a social networking and identity service that is owned and operated by Google. *LiveJournal* is a social networking service where users can keep a blog, journal, or diary. *YouTube* is a popular video-sharing service/website where users can share their videos with friends, family, and the world. *Orkut* is a social networking service which is designed to help users meet new and old friends and maintain existing relationships. *Pokec* is one of the most popular online social networking services in Slovakia. *Facebook* is one of the most popular online social networking services in the world, where users can create their profiles, add other as "friends", exchange messages, post status updates and photos, share videos, and receive notifications when others update their profiles. *Flickr* is an image and video hosting website/social network service. *Foursquare* is a local search and discovery mobile service which provides a personalized local search experience for its users. *Twitter* is an online social networking service that enables users to send and read short 140-character messages called "tweets".

• **Location-based Social Network (LSN) Graph.** *Gowalla* is a location-based social networking service, where users are able to check in at "Spots" in their local vicinity, either through a dedicated mobile application or through the mobile website.

• **Collaboration (Collab.) Graph.** *AstroPh* is from the e-print arXiv and covers scientific collaborations between authors-papers submitted to the Astro Physics category.

• **Email Graph.** *Enron* is an email graph that was originally made public and posted to the web by the Federal Energy Regulatory Commission during its investigation. *EuAll* is an email graph that was generated using the email data from a large European research institution.

• **Autonomous System (AS) Graph.** *Skitter* is an Internet topology graph.

• **P2P Graph.** *Gnutella* is the network topology graph of the P2P network Gnuttella.

B. Evaluation Methodology

To ensure that our results can be fairly compared with those in [5], [8], we use the same evaluation setup as in [5], [8], i.e., we do not preprocess the employed datasets, e.g., removing low-degree users. Furthermore, as in [5], [8], we assume $s_a = s_u = s$ for convenience. Note that this assumption does not limit the generality of our evaluation. All the evaluations can be extended to the case of $s_a \neq s_u$ directly.

In our evaluation, we mainly focus on evaluating the impacts of θ and s on the *inherent* de-anonymizability of each dataset using our quantifications in Section IV and Section V, i.e., how theoretically de-anonymizable the users of each dataset are with respect to their structural properties. Let Ξ be a DA attack without any computational limitation and Ξ de-anonymizes G^a leveraging G^u by minimizing the *edge difference function* Ψ_σ defined in Section III. Therefore, the output DA scheme of Ξ is the mapping scheme σ^* such that $\sigma^* = \arg\min_{\sigma} \Psi_\sigma$, i.e., σ^* has the minimum edge difference among all the possible DA schemes[3]. Then, statistically, each evaluation result in this section can be considered as the performance lower bound of Ξ under a specific setting of θ and s.

C. Seed-based RD Evaluation

In this subsection, we evaluate the seed-based RD of the 15 datasets. The number of available seeds is assumed to be 50, i.e., $\Lambda = 50$. To simplify the evaluation process, we take the users with the top-50 degrees in each dataset as the seeds.

1) RD versus θ: When $s = 0.6$ and $\Lambda = 50$, we evaluate the seed-based RD of each dataset with respect to different θ as shown in Table II, where each value indicates the percentage of users that can be successfully de-anonymized. From Table II, we have the following observations.

When θ increases, the seed-based RD of each dataset decreases. For instance, when $\theta = 0.05$ (i.e., the target DA users are the top-5% users with respect to degree), 100% of target users of Google+ can be successfully de-anonymized, while when $\theta = 0.45$ (the target users now are the top-45% users with respect to degree), the successfully de-anonymizable target users of Google+ are decreased to 60.1%. The reason is that when θ increases, more users with relatively low degrees become target users of DA. However, less structural information is available for the the new considered low-degree users, and thus the RD of each dataset decreases.

Generally, the datasets with higher average degree (i.e., $\overline{d}$) and graph density (i.e., ρ) are more de-anonymizable than the datasets with lower average degree and graph density. For instance, as shown in Table I, Google+ has a similar size with LiveJournal. However, Google+ has a higher average degree/graph density than LiveJournal. According to Table II, Google+ is more de-anonymizable than LiveJournal, e.g., when $\theta = 0.3$, 94.6% target users of Google+ can be de-anonymized while 57.2% target users of LiveJournal can be de-anonymized. This is because Google+ has a larger $\overline{d}/\rho$, and thus more structural information is available to uniquely and correctly distinguish the users in Google+ compared to LiveJournal.

2) RD versus s: When $\theta = 0.5$, the seed-based de-anonymizability of the 15 datasets under different s is shown in Table III, where each value indicates the percentage of target users that can be successfully de-anonymized. From Table III, we have the following observations.

When s increases, more users can be successfully de-anonymized. For instance, when $s = 0.4$, 23.7% of the target users of Google+ can be successfully de-anonymized and when $s = 0.7$, the ratio of target users that can be

[3]Given G^a and G^u, it is intuitively that there are at most $n!$ possible DA schemes, i.e., at most $n!$ mapping schemes from V^a to V^u. Then, if there is no computational limitation on Ξ, Ξ can minimize Ψ_σ by examining all the $n!$ possible DA schemes.

TABLE II
SEED-BASED DE-ANONYMIZABILITY ANALYSIS ($s = .6, \Lambda = 50$).

$\theta =$	.05	.1	.15	.2	.25	.3	.35	.4	.45
Google+	100%	100%	100%	100%	100%	94.6%	79.8%	68.7%	60.1%
LiveJournal	100%	100%	100%	88.7%	69.8%	57.2%	48.2%	41.4%	36.1%
YouTube	100%	98.1%	64.5%	47.6%	37.5%	30.8%	26.0%	22.3%	19.5%
Orkut	100%	100%	100%	100%	100%	100%	100%	98.5%	86.0%
Pokec	100%	100%	100%	100%	87.3%	71.4%	60.0%	51.5%	44.9%
Facebook	100%	100%	100%	100%	99.6%	84.6%	71.2%	61.2%	53.3%
Flickr	100%	100%	100%	100%	100%	100%	100%	100%	100%
Foursquare	100%	100%	99.9%	76.4%	60.2%	49.4%	41.7%	35.8%	31.3%
Twitter	100%	100%	100%	100%	100%	100%	100%	93.7%	81.9%
Gowalla	100%	100%	93.4%	69.0%	54.3%	44.5%	37.5%	32.2%	28.1%
AstroPh	99.9%	99.9%	99.9%	99.7%	95.3%	78.7%	66.2%	56.8%	49.5%
Enron	99.9%	99.9%	99.8%	84.8%	66.8%	54.8%	46.2%	39.8%	34.8%
EuAll	100%	79.0%	52.0%	38.5%	30.3%	24.9%	21.0%	18.1%	15.8%
Skitter	100%	100%	100%	76.6%	60.4%	49.5%	41.7%	35.9%	31.3%
Gnutella	99.5%	71.7%	48.6%	36.9%	29.9%	25.3%	22.0%	19.4%	17.5%

TABLE III
SEED-BASED DE-ANONYMIZABILITY ANALYSIS ($\theta = .5, \Lambda = 50$).

$s =$	.35	.4	.45	.5	.55	.6	.65	.7	.75
Google+	18.4%	23.7%	29.7%	36.6%	44.4%	53.2%	63.0%	74.0%	86.3%
LiveJournal	9.9%	13.2%	17.0%	21.3%	26.3%	31.9%	38.2%	45.1%	52.9%
YouTube	6.1%	7.9%	9.8%	12.0%	14.5%	17.2%	20.3%	23.6%	27.2%
Orkut	21.5%	29.1%	38.2%	48.9%	61.5%	76.0%	92.5%	100%	100%
Pokec	10.5%	14.5%	19.4%	25.2%	31.8%	39.5%	48.3%	58.1%	69.2%
Facebook	12.3%	17.4%	23.3%	30.2%	38.1%	47.1%	57.1%	68.2%	80.5%
Flickr	46.7%	62.3%	80.2%	99.5%	100%	100%	100%	100%	100%
Foursquare	9.3%	12.2%	15.4%	19.1%	23.1%	27.7%	32.6%	38.0%	43.8%
Twitter	23.6%	31.0%	39.5%	49.1%	60.1%	72.5%	86.4%	100%	100%
Gowalla	7.9%	10.5%	13.4%	16.8%	20.6%	24.8%	29.6%	34.9%	40.8%
AstroPh	9.6%	14.8%	20.5%	27.2%	34.9%	43.6%	53.4%	64.4%	76.5%
Enron	9.7%	13.1%	16.9%	21.0%	25.6%	30.7%	36.3%	42.5%	49.4%
EuAll	5.4%	6.8%	8.4%	10.1%	11.9%	14.0%	16.3%	18.7%	21.4%
Skitter	9.5%	12.3%	15.5%	19.1%	23.2%	27.7%	32.8%	38.4%	44.7%
Gnutella	7.2%	8.3%	9.6%	11.3%	13.4%	16.0%	19.0%	22.6%	26.8%

TABLE IV
SEED-FREE DE-ANONYMIZABILITY ANALYSIS ($s = .6$).

$\theta =$	.05	.1	.15	.2	.25	.3	.35	.4	.45
Google+	100%	100%	100%	87.4%	75.8%	67.7%	61.6%	56.8%	52.9%
LiveJournal	94.0%	63.6%	51.3%	44.2%	39.5%	36.0%	33.4%	31.2%	29.5%
YouTube	51.7%	34.8%	27.9%	23.9%	21.2%	19.3%	17.8%	16.6%	15.6%
Orkut	100%	100%	100%	100%	100%	100%	91.9%	85.0%	79.4%
Pokec	99.9%	73.3%	60.6%	53.0%	47.9%	44.1%	41.1%	38.7%	36.7%
Facebook	99.7%	98.3%	81.3%	69.8%	62.1%	56.4%	52.1%	48.6%	45.8%
Flickr	100%	100%	100%	100%	100%	100%	100%	100%	100%
Foursquare	89.5%	59.0%	46.8%	39.9%	35.4%	32.0%	29.4%	27.4%	25.7%
Twitter	100%	100%	100%	100%	100%	99.1%	89.7%	82.2%	76.3%
Gowalla	73.2%	49.6%	39.9%	34.3%	30.6%	27.9%	25.8%	24.1%	22.7%
AstroPh	96.3%	85.0%	70.7%	61.7%	55.5%	50.8%	47.2%	44.3%	41.9%
Enron	97.1%	68.6%	54.1%	45.9%	40.4%	36.5%	33.5%	31.1%	29.1%
EuAll	46.4%	30.4%	23.9%	20.2%	17.8%	16.1%	14.7%	13.7%	12.8%
Skitter	91.1%	59.3%	46.8%	39.8%	35.2%	31.9%	29.3%	27.3%	25.7%
Gnutella	18.7%	17.4%	16.4%	15.7%	15.1%	14.7%	14.3%	14.0%	13.8%

successfully de-anonymized is increased to 74%. This is because a large s implies more common edges are shared by G^a (the anonymized graph) and G^u (the auxiliary graph), i.e., more structural similarity between G^a and G^u. Hence, more users in G^a can be successfully de-anonymized leveraging the increased structural similarity between G^a and G^u.

Even if s is small, a significant number of users within each dataset can still be successfully de-anonymized, e.g., 18.4% target users of Google+ and 46.7% target users of Flickr can be successfully de-anonymized when $s = 0.35$. This demonstrates that structure-based DA is very powerful. Therefore, in addition to protect the semantic information carried by graph data, the data's structural information is also important and deserves dedicated consideration/protection.

Similar to analyze Table II, if a dataset has higher average degree and/or graph density, it is more vulnerable to structure-based DA attacks, e.g., Flickr is more vulnerable than Google+. The reason is the same as analyzed before: higher average degree and/or graph density imply more structural information can be used for conducting successful DA.

TABLE V
SEED-FREE DE-ANONYMIZABILITY ANALYSIS ($\theta = .5$).

$s =$	.35	.4	.45	.5	.55	.6	.65	.7	.75
Google+	15.8%	20.6%	26.3%	32.9%	40.7%	49.8%	60.3%	72.6%	86.9%
LiveJournal	8.1%	10.9%	14.2%	18.1%	22.7%	28.0%	34.1%	41.2%	49.2%
YouTube	5.1%	6.5%	8.2%	10.2%	12.3%	14.8%	17.5%	20.6%	24.0%
Orkut	18.0%	25.0%	33.7%	44.7%	58.2%	74.8%	94.8%	100%	100%
Pokec	8.3%	11.7%	16.0%	21.2%	27.5%	35.0%	44.0%	54.6%	67.1%
Facebook	9.8%	14.2%	19.7%	26.2%	34.1%	43.4%	54.2%	66.8%	81.3%
Flickr	43.1%	60.1%	80.8%	99.9%	100%	100%	100%	100%	100%
Foursquare	7.7%	10.2%	13.0%	16.4%	20.1%	24.3%	29.2%	34.5%	40.5%
Twitter	20.5%	27.4%	35.7%	45.7%	57.5%	71.3%	87.6%	100%	100%
Gowalla	6.5%	8.7%	11.2%	14.2%	17.6%	21.6%	26.1%	31.2%	37.1%
AstroPh	6.5%	11.6%	16.9%	23.2%	30.8%	39.8%	50.4%	62.8%	77.0%
Enron	7.9%	10.9%	14.4%	18.2%	22.6%	27.5%	33.1%	39.4%	46.5%
EuAll	4.5%	5.8%	7.1%	8.6%	10.3%	12.1%	14.1%	16.3%	18.8%
Skitter	7.9%	10.3%	13.1%	16.4%	20.1%	24.3%	29.1%	34.7%	41.0%
Gnutella	6.5%	7.4%	8.5%	9.8%	11.5%	13.6%	16.0%	19.1%	22.9%

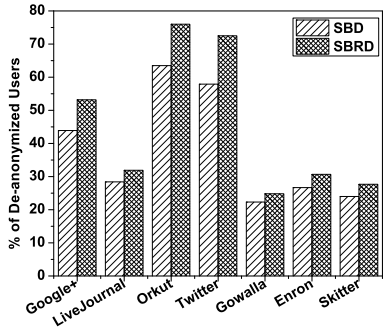
D. Seed-free RD Evaluation

1) RD versus θ: When we fix $s = 0.6$, the seed-free RD of the 15 datasets given different θ is shown in Table IV. From Table IV, we have the following observations.

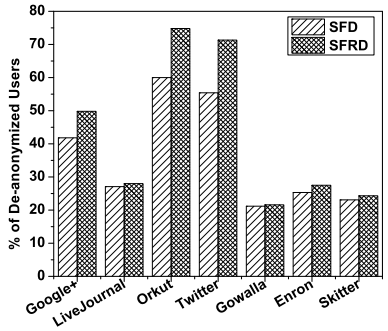
Generally, all the datasets are completely or partially de-anonymizable even if there is no seed information available, i.e., the considered graph datasets are completely/partially de-anonymizable based only on their structural information. For instance, when $\theta = 0.3$, 94.6%, 84.6%, and 100% target users of Google+, Facebook, and Twitter can be successfully de-anonymized, respectively. This is because, as shown in our de-anonymizability quantification, the structural information associated with users (especially high-degree users) is sufficient to uniquely distinguish them with a high probability. Therefore, our seed-free RD quantification provides the theoretical foundation for the success of emerging *seed-free* DA attacks, e.g., Bayesian-based attack [4] and optimization-based attack [5].

When θ increases, the target users of each dataset (except for Flickr) become less de-anonymizable. For instance, 69.8% target users of Facebook are de-anonymizable when $\theta = 0.2$ while 48.6% target users of Facebook are de-anonymizable when $\theta = 0.4$. The reason is the same as analyzed in the seed-based DA scenario: with the increase of θ, more relatively low-degree users become to target DA users. However, less structural information can be employed to de-anonymize the relatively low-degree users, followed by the decrease of the percentage of the de-anonymizable users in each dataset.

When comparing the results in Table IV (the seed-free scenario) with the results in Table II (the seed-based scenario), we find that each dataset is more de-anonymizable in the seed-based scenario than that in the seed-free scenario, which is consistent with the intuition given the same θ. For instance, when $\theta = 0.3$, 84.6% target users of Facebook can be successfully de-anonymized in the seed-based DA scenario while 56.4% target users of Facebook can be successfully de-anonymized in the seed-free DA scenario. Therefore, although the carried structural information of graph data is sufficient to conduct large-scale successful DA, we conclude that the available seed information can improve the DA performance.



(a) SBRD vs SBD ($\Lambda = 50$)



(b) SFRD vs CCS'14

Fig. 2. Comparisons with state-of-the-art quantification techniques [5], [8].

Similar to the seed-based DA scenario, if a dataset has higher average degree and/or graph density, it is also more de-anonymizable in the seed-free DA scenario. Again, the reason is due to the carried richer structural information.

2) RD versus s: When fixing $\theta = 0.5$, the seed-free RD of the 15 datasets with respect to different s is shown in Table V, from which we have the following observations.

Similar to the seed-based DA scenario, with the increase of s, all the datasets become more de-anonymizable. For instance, when $s = 0.35$, 15.8% target users of Google+ can be successfully de-anonymized while when $s = 0.75$, 86.9% target users of Google+ can be successfully de-anonymized. The reason is also the same as in the seed-based DA scenario:

a large s implies G^a is more structurally similar to G^u, and thus more users in G^a can be successfully de-anonymized leveraging G^u.

When comparing Table V (the seed-free scenario) with Table III (the seed-based scenario), we find that each dataset is more de-anonymizable when having seed information available. For instance, when $s = 0.5$, 49.1% target users of Twitter can be successfully de-anonymized in the seed-based DA scenario while 45.7% target users of Twitter can be successfully de-anonymized in the seed-free DA scenario. Again, the reason is straightforward: the available seed users can provide more accurate auxiliary information in the DA.

Due to the same reason as analyzed before, the datasets with the higher average degree and/or graph density are more de-anonymizable than the datasets with lower average degree and/or graph density.

E. Comparisons with State-of-the-Art Quantifications

In this subsection, we experimentally compare our Seed-Based RD quantification (SBRD) with the state-of-the-art seed-based de-anonymizability quantification technique in [8], denoted by *SBD*, and compare our Seed-Free RD quantification (SFRD) with the state-of-the-art seed-free de-anonymizability quantification technique in [5], denoted by *SFD*. Since the common datasets employed in [5], [8], and this paper are Google+, LiveJournal, Orkut, Twitter, Gowalla, Enron, and Skitter, we employ these 7 datasets to conduct the comparative evaluation study. When comparing SBD and SBRD, we select 50 top-degree users from each dataset to serve as seeds. Furthermore, in all the evaluation, we set $s = s_a = s_u = 0.6$ for convenience.

Leveraging the 7 datasets, we show the seed-based de-anonymizability evaluation results of SBD and SBRD in Fig.2 (a) and show the seed-free de-anonymizability evaluation results of SFD and SFRD in Fig.2 (b). From Fig.2, we make the following observations.

When comparing SBD with SBRD (Fig.2 (a)), we find that all the 7 datasets are actually more de-anonymizable under SBRD that under SBD. Therefore, the results implies the seed-based DA conditions obtained in this paper are more accurate/tighter than that in [8]. The reason is because instead of considering all the users structurally equivalent, we adaptively quantify the de-anonymizability of users according to their structural importance.

Similarly, when comparing SFD and SFRD (Fig.2 (b)), all the datasets are more de-anonymizable under SFRD than under SFD, which implies our seed-free RD quantification technique is more accurate than that in [5]. The reason is the same as in the seed-based de-anonymizability quantification scenario: adaptively considering users' structural importance enables us to quantify the de-anonymizability of graph data in a more accurate manner.

Remark. For sparse graph datasets (e.g., LiveJournal), the improvements of our relative quantifications are not very significant compared with state-of-the-art quantifications. This is because less structural information is carried by sparse graphs, i.e., they are *inherently* more resistant to structure-based DA attacks. The most notable contribution of our RD quantifications is that *we fundamentally/theoretically provide more accurate/tighter seed-based and seed-free DA bounds for anonymized graph data* than state-of-the-art quantifications.

VII. Conclusion

In this paper, we study the structural importance-aware RD quantification problem for graph data. Specifically, we quantify both the seed-based and the seed-free RD of anonymized graph data for both perfect DA (de-anonymizing all the target users) and partial de-anonymizaiton (tolerating some DA error) under a general graph model. In our quantification, instead of treating all the users in a graph dataset as structurally equivalent, we adaptively quantify their actual de-anonymizability in terms of their structural importance. Leveraging 15 real world graph datasets, we evaluate our relative quantifications and compare them with state-of-the-art seed-based and seed-free quantification techniques. The results demonstrate that our structural importance-aware relative quantifications are more accurate when measuring graph data's real de-anonymizability.

Acknowledgment

This work was supported in part by NSF awards number CNS-1409415 and CNS-1423139.

References

[1] A. Narayanan and V. Shmatikov. De-anonymizing social networks. *S&P*, 2009.
[2] S. Ji, W. Li, M. Srivatsa, J. He, and R. Beyah. Structure based data de-anonymization of social networks and mobility traces. *ISC*, 2014.
[3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. *WWW*, 2007.
[4] P. Pedarsani, D. R. Figueiredo, and M. Grossglauser. A bayesian method for matching two similar graphs without seeds. *Allerton*, 2013.
[5] S. Ji, W. Li, M. Srivatsa, and R. Beyah. Structural data de-anonymization: Quantification, practice, and implications. *CCS*, 2014.
[6] M. Srivatsa and M. Hicks. Deanonymizing mobility traces: Using social networks as a side-channel. *CCS*, 2012.
[7] 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.
[8] 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.
[9] P. Pedarsani and M. Grossglauser. On the privacy of anonymized networks. *KDD*, 2011.
[10] L. Yartseva and M. Grossglauser. On the performance of percolation graph matching. *COSN*, 2013.
[11] N. Korula and S. Lattanzi. An efficient reconciliation algorithm for social networks. *VLDB*, 2014.
[12] M. E. J. Newman. Networks: An introduction. *Oxford University Press*, 2010.
[13] A. Narayanan, E. Shi, and B. Rubinstein. Link prediction by de-anonymization: How we won the kaggle social network challenge. *IJCNN*, 2011.
[14] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn. Community-enhanced de-anonymization of online social networks. *CCS*, 2014.
[15] P. Mittal, C. Papamanthou, and D. Song. Preserving link privacy in social network based systems. *NDSS 2013*.
[16] Stanford Large Network Dataset Collection. http://snap.stanford.edu/data/index.html.
[17] ASU Social Computing Data Repository. http://socialcomputing.asu.edu/pages/datasets.