Structural Data De-anonymization: Quantification, Practice, and Implications

Shouling Ji
School of Electrical and Computer Engineering
Georgia Institute of Technology
sji@gatech.edu

Weiqing Li
School of Electrical and Computer Engineering
Georgia Institute of Technology
wli64@gatech.edu

Mudhakar Srivatsa
IBM T. J. Watson Research Center
msrivats@us.ibm.com

Raheem Beyah
School of Electrical and Computer Engineering
Georgia Institute of Technology
rbeyah@ece.gatech.edu

ABSTRACT

In this paper, we study the quantification, practice, and implications of structural data (e.g., social data, mobility traces) De-Anonymization (DA). First, we address several open problems in structural data DA by quantifying *perfect* and $(1-\epsilon)$*-perfect structural data DA*, where ϵ is the *error* tolerated by a DA scheme. To the best of our knowledge, this is the first work on quantifying structural data DA under a *general data model*, which closes the gap between structural data DA practice and theory. Second, we conduct the first large-scale study on the de-anonymizability of 26 real world structural datasets, including Social Networks (SNs), Collaborations Networks, Communication Networks, Autonomous Systems, and Peer-to-Peer networks. We also quantitatively show the conditions for perfect and $(1-\epsilon)$-perfect DA of the 26 datasets. Third, following our quantification, we design a practical and novel *single-phase cold start Optimization based DA* (ODA) algorithm. Experimental analysis of ODA shows that about $77.7\%-83.3\%$ of the users in Gowalla (.2M users and 1M edges) and $86.9\%-95.5\%$ of the users in Google+ (4.7M users and 90.8M edges) are de-anonymizable in different scenarios, which implies optimization based DA is implementable and powerful in practice. Finally, we discuss the implications of our DA quantification and ODA and provide some general suggestions for future *secure data publishing*.

Categories and Subject Descriptors

C.2.0 [**General**]: Security and protection; H.4 [**Information Systems Applications**]: Miscellaneous; G.3 [**Probability and Statistics**]: Stochastic processes

General Terms

Security, Privacy, Theory, Management

CCS'14, November 3–7, 2014, Scottsdale, Arizona, USA.

http://dx.doi.org/10.1145/2660267.2660278.

Keywords

De-anonymization; structural data; quantification; evaluation; social networks; mobility traces

1. INTRODUCTION

Nowadays, a large amount of data generated by computer networks and services have a graph structure, which is referred to as *structural data*. For instance, it is straightforward to model Social Networks (SNs), network topologies, etc. by graphs [2][5][6][26]. Additionally, mobility traces (e.g., WiFi contacts, Instant Message contacts) can also be modeled as graphs (structural data) [3]. Even general spatiotemporal data (mobility traces) with the classical (*latitude*, *longitude*, *timestamp*) format can be converted to structural data by applying sophisticated techniques [27]. Since these structural data have huge commercial value to businesses and potentially significant impacts to society [28][29], the security and privacy issues that arise during data release to the public, sharing with commercial partners, and/or transferring to third parties are attracting increasing interest [1][2][3].

Currently, to protect structural data's privacy, the most common technique used is to anonymize data by removing the "*Personally Identifiable Information* (PII)" before releasing data. Unfortunately, this naive method is shown to be vulnerable to many *De-Anonymization* (DA) *attacks* [6][7][8]. Latterly, some sophisticated anonymization schemes to protect structural data privacy, e.g., k-anonymity and its variants [6][7][8], were designed[1]. They can protect the privacy of structural data to some extent. However, they are susceptible to emerging *structure based DA attacks* due to the limitations of the schemes (e.g., they are syntactic properties based) and the rich amount of information available to adversaries [1][2][3] (see the detailed analysis in Section 2).

In structure based DA attacks, some auxiliary data (graphs) are employed to break the privacy of anonymized structural data based only on the structural information. The fact that the auxiliary data may come from either the same or a different domain/context with the anonymized data makes

[1]Also, *differential privacy* [9] was developed to protect the privacy of *interactive data release*. However, it cannot defend against *structural data DA attacks* which breach the privacy of *non-interactive data release* [2][3][8][9].

the attack powerful, e.g., using Flickr to de-anonymize Twitter [2], using Facebook to de-anonymize WiFi mobility traces [3]. Furthermore, the wide availability of auxiliary data makes the attack applicable and practical [2][3].

Structure based DA attacks were initially presented in [1], where Backstrom et al. designed both active and passive attacks to break the privacy of SN users. However, since the attacks in [1] leverage the success of a "sybil" attack before actual anonymized data publication, they are less practical. Later, Narayanan and Shmatikov designed a new structure based DA attack in [2], which successfully de-anonymized a large scale directed social network by applying several heuristics such as eccentricity, edge directionality, reverse match. In [3], Srivatsa and Hicks demonstrated that the privacy of three kinds of mobility traces can be compromised by structure based DA attacks. However, the attacks presented in [3] are only suitable for small datasets due to the computational infeasibility of finding a proper landmark mappings for large datasets. Furthermore, each of the aforementioned attacks consist of two phases: a *landmark identification phase* and a *DA propagation phase*.

Although we already have some successful structure based DA techniques [1][2][3], we do not have any *rigorous theoretical analysis under a general model* yet that explains why structure based DA attacks work. In [5], Pedarsani and Grossglauser quantified the privacy of anonymized structural data under the *Erdös-Rényi (ER) random graph model* $G(n,p)$ (every edge exits with identical probability p). However, this quantification is not suitable in practice since most, if not all, observed real world structural data (e.g., SNs, collaboration networks [21][22][26]) do not follow the ER model. Actually, they may follow the *power-law model*, *exponential model*, etc. [21][22][26]. Therefore, under a practical *general data model*, there are still some open questions in DA research, including: *why can structural data be de-anonymized*? *what are the conditions for successful structural data DA*? and *what portion of users can be de-anonymized in a structural dataset*? To close the practice-theory gap, we study the *quantification*, *practice*, and *implications* of structural data DA in this paper. Specifically, our contributions are as follows.

- To the best of our knowledge, this is the first work on quantifying structural data DA under a general data model. In our quantification, we answer several fundamental open questions: why can structural data be de-anonymized based only on the topological information (the inherent reason for the success of existing structure based DA practices)? what are the conditions for *perfect* and $(1-\epsilon)$*-perfect DA*, where ϵ is the *error* tolerated by a DA scheme? what portion of users can be de-anonymized in a structural dataset? Thus, we close the gap between structural data DA practice and theory.

- We conduct the first large-scale study on the de-anonymizability of 26 real world structural datasets, including SNs, location based mobility traces and SNs, collaboration networks, communication networks, autonomous systems, peer-to-peer networks, etc. Based on our study, we find that *all* the structural datasets that we considered are perfectly or partially de-anonymizable. We also quantitatively show the conditions for perfect and $(1-\epsilon)$-perfect DA and what portion of users can be de-anonymized for the 26 datasets.

- Following our quantification, we present a novel *Optimization based DA* (ODA) attack. Different from existing structure based DA attacks [1][2][3], ODA is a *single-phase cold start* algorithm without any requirement on priori knowledge, e.g., landmark mappings. We also examine ODA on real datasets Gowalla (.2M users and 1M edges) and Google+ (4.7M users and 90.8M edges). The results demonstrate that about $77.7\% - 83.3\%$ of the users in Gowalla and $86.9\% - 95.5\%$ of the users in Google+ are de-anonymizable, which illustrates that optimization based DA is implementable and powerful in practice.

- Finally, we discuss some implications of this work according to our structural DA quantification and the ODA attack. We further provide some general suggestions for future *secure data publishing*.

The rest of this paper is organized as follows. In Section 2, we summarize the related work. In Section 3, we give the data and attack models. In Section 4, we theoretically quantify perfect and $(1-\epsilon)$-perfect DA attacks under a general model, followed by a large-scale evaluation in Section 5. In Section 6, we present a novel optimization based DA attack. We discuss the implications in Section 7. The paper is concluded and future work is addressed in Section 8.

2. RELATED WORK

In this section, we first briefly survey the state-of-the-art advances of security issues related to structural data. Subsequently, we discuss the status quo on structural data anonymization and DA followed by remarking the characteristics that distinguish this paper from existing works.

2.1 State-of-the-Art Advances

Recently, the security and privacy issues related to structural data have attracted the interest of many researchers [1]-[20]. In [10], Korolova et al. presented an attack on link privacy in SNs. Another attack using link-based and group-based classification to study privacy implications in SNs is discussed in [11] by Zheleva and Getoor. Based on four previously unrecognized implicit identifiers, Pang et al. developed an automated procedure to identify users in 802.11 traces [12]. In [13], Backstrom proposed an algorithm to predict users' location using SN information. On the other hand, multiple strategies have been developed to protect people's privacy in SN systems and related applications, e.g., *pseudonym abstraction* [14], *decentralized protocols for anonymous communications* [15], *guaranteed data lifetime* [16], *compromised accounts detection* [17], *privacy preserving SN applications* [18]. In addition, location based services, especially those in smartphone-based SN applications, have created big commercial benefits. However, on the other hand, the publicly availability of users' mobility trace data causes a potentially serious threat to users' privacy and even themselves [3][19][20].

2.2 Structural Data Anonymization and DA

2.2.1 Anonymization Schemes

To protect the privacy of structural data, the most common method is removing the PII [1][2][3]. However, this widely used naive solution is proven to be vulnerable to many DA attacks [6][7][8]. Later, researchers proposed some sophisticated data anonymization solutions, e.g., k-anonymity and its many variants [6][7][8]. These solutions do work to protect users' privacy against semantics based DA attacks to some extent. However, according to recent empirical studies [2][19], these solutions fail against emerging structure based DA attacks. Some of the reasons are as follows. First, the adversaries may obtain much richer auxiliary information through multiple channels, e.g., data mining, advertising, third-party applications, data aggregation. [1][2][3][4]. Second, the k-anonymity scheme is applicable to low-average-degree datasets. Nevertheless, many datasets, e.g., Google+ [25], Facebook [26], tend to have a large average degree and still increasing. Finally, the k-anonymity idea lies on data's syntactic property, which may not work on protecting data privacy even if this property is satisfied [2].

2.2.2 DA Schemes

Several successful DA attacks on structural data were proposed recently, including structure based schemes [1][2][3] and semantics based schemes [4]. In [4], Wondracek et al. designed a DA attack to SN users based on the *group membership* information. To implement this attack, the adversary should collect enough group membership information (semantics information) by "*history stealing*" on browsers, and then try to uniquely identify a user.

Instead of leveraging semantic information, we focus on structure based DA attacks in this paper. In [1], Backstrom et al. introduced the structure based DA attacks by designing both active and passive attacks in SNs. Since the designed attacks in [1] leverage a "sybil" attack before the actual anonymized data release, they are not scalable as SNs increase in size [2]. Later, Narayanan and Shmatikov in [2] designed a two-phase heuristics based DA attack against *large scale directed SNs*. Through experiments on real datasets, they demonstrated the feasibility of large scale DA in terms of structure information. In [3], Srivatsa and Hicks presented three DA attacks to mobility traces, which are all two-phase schemes. In all the three attacks, to achieve high DA accuracy, the DA propagation (the second phase) must be repeated for all $k!$ possible landmark mappings (k is the number of landmarks), which is very time-consuming. For instance, to de-anonymize a small dataset having 125 users with 5 landmarks, the three attacks take 6.7 hours, 6.2 hours, and 0.5 hours, respectively. Furthermore, the attacks in [3] are not suitable for large scale DA since in that scenario more landmarks are required ($k > 30$, [2]), while when $k \geq 20$, the attacks in [3] are computationally infeasible.

2.2.3 Remarks

In this paper, we study the *quantification*, *practice*, and *implications* of structural data DA. The main aspects distinguishing this paper from existing works are as follows. (i) To the best of our knowledge, this is the first work that quantifies structural data DA under a general model. By our quantification, we answered several open questions, e.g., why can structural data be de-anonymized? what are the conditions for structural data DA? what portion of users can be de-anonymized? and thus we bridge the gap between structural data DA practice and theoretical quantification. (ii) Following our theoretical quantification, we conduct a large scale study on 26 real world structural datasets. We also conduct comprehensive and in-depth analysis on the evaluation results. (iii) Following our quantification, we propose a novel single-phase optimization based DA algorithm (ODA). ODA is a cold start algorithm without any requirement on priori knowledge. By conducting experiments on large scale real datasets, we demonstrate the effectiveness of ODA.

3. SYSTEM MODEL

In this paper, we focus on quantifying the DA attack (vulnerability) on anonymized structural data, which could be social data released by SN operators, (e.g., Google+ [25]) and/or mobility data generated by mobile devices (e.g., classical longitude-latitude spatiotemporal traces [26][27]).

3.1 Data Model

It is straightforward to model social data using graphs, where nodes represent users and edges indicate the social relationships (*friendship*, *contact*, *following*) among users. Mobility data generated by users (users' devices) can also be modeled by contact graphs [3][27]. Furthermore, it has been shown that a contact graph derived from mobility data has strong correlation with the social graph of the same group of users that generated them [3][27]. Therefore, we model the anonymized structural data by a graph $G^a = (V^a, E^a)$, where $V^a = \{i | i \text{ is an } \textit{anonymized} \text{ user}\}$ is the user set and $E^a = \{e^a_{i,j} | \text{ there is a relationship between } i \in V^a \text{ and } j \in V^a\}$ is the edge/relationship set. In reality, it is possible that a structural dataset corresponds to a directed graph, e.g., Twitter. However, for simplicity and without loss of generality, we assume G^a as an undirected graph. Note that, the designed algorithm in this paper can be extended to the directed scenario directly. For $i \in V^a$, its neighborhood is defined as $N^a_i = \{j | \exists e^a_{i,j} \in E^a\}$ and we denote the cardinality of N^a_i as $|N^a_i|$, i.e., the degree of i.

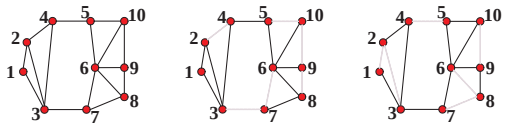
The auxiliary data is also assumed to be structural data, e.g., a SN compromising users overlapped with that in G^a [2][3]. Furthermore, the auxiliary data is easily obtainable by multiple means such as academic and government data mining, advertising, third-party applications, data aggregation, online crawling, etc. Successful examples can be found in [2][3][4][27]. Consequently, the auxiliary data is also modeled by a graph $G^u = (V^u, E^u)$, where $V^u = \{i \text{ is a } \textit{known} \text{ user}\}$ and $E^u = \{e^u_{i,j} | \text{ there is a relationship between } i \in V^u \text{ and } j \in V^u\}$. Similarly, the neighborhood of $i \in V^u$ is defined as $N^u_i = \{j | \exists e^u_{i,j} \in E^u\}$.

3.2 DA Attack

Given G^a and G^u, a DA attack can be formally defined as a *mapping*: $\sigma : V^a \rightarrow V^u$. For $\forall i \in V^a$, its mapping under σ is $\sigma(i) \in V^u \cup \{\perp\}$, where $\perp$ is a special *not existing indicator*. Similarly, for $\forall e^a_{i,j} \in E^a$, $\sigma(e^a_{i,j}) = e^u_{\sigma(i),\sigma(j)} \in E^u \cup \{\perp\}$. Under σ, a successful DA on $i \in V^a$ is defined as $\sigma(i) = i'$, if $i' \in V^u$ and i and i' correspond to the same user; or $\sigma(i) = \perp$, otherwise. For other cases, the DA on i fails. Consequently, the objective of a DA attack is to successfully de-anonymize as many users in V^a as possible.

4. DA QUANTIFICATION

In this section, given G^a and G^u, we quantify a DA attack under an *arbitrary graph distribution* in multiple scenarios.



(a) G (b) G^a (c) G^u

Figure 1: Edge/relationship projection. Only black edges appear in G^a/G^u.

Note that, our quantification aims to provide a *theoretical foundation* for understanding the success of recent heuristic structure-based DA practices [2][3].

4.1 Preliminaries

To make the quantification and proof tractable and convenient, we make some assumptions and definitions. First, we assume $V^a = V^u$, i.e., the auxiliary data and the anonymized data are corresponding to the same group of users [2][3][5]. This does not mean that we know any priori correct mapping from V^a to V^u. Furthermore, this assumption is reasonable since one cannot be expected to use G^u to de-anonymize G^a if they correspond to different groups of users. It is possible that the auxiliary data only has some overlap with the anonymized data instead of corresponding to the exactly same group of users. This fact does not limit our theoretical analysis since we can either (i) apply the quantification to the overlapping part, or (ii) redefine $V^a_{new} = V^a \cup (V^u \setminus V^a)$ and $V^u_{new} = V^u \cup (V^a \setminus V^u)$, i.e. adding the non-overlapped users to V^a and V^u respectively as isolated users (with degree 0), and apply the analysis to $G^a = (V^a_{new}, E^a)$ and $G^u = (V^u_{new}, E^u)$. Without of causing any confusion, we assume $V^a = V^u$ in the rest of this section.

Second, similar to [5], for the users in V^a (or, V^u), we assume that there exists a conceptual underlying graph $G = (V, E)$ with $V = V^a = V^u$ and E consisting of the true relationships among users in V. Consequently, G^a and G^u can be viewed as the physically observable *projections* of G on particular relationships, e.g., "circle" relationship on Google+, "co-occurrence" relationship in Gowalla. The projection from G to G^a is characterized by an *edge/relationship projection process* [5]: (i) $V^a = V$; and (ii) $\forall e_{i,j} \in E$, $e_{i,j}$ is appeared in E^a with probability p_a, i.e., $\Pr(e_{i,j} \in E^a | e_{i,j} \in E) = p_a$. Similarly, the projection from G to G^u can be characterized by another *edge/relationship projection process* with probability p_u. For instance, we show a projection from G to G^a/G^u in Fig. 1. Furthermore, we assume both projection processes are *independent and identically distributed* (*i.i.d.*). Note that, (i) although the assumption on the existence of a conceptual underlying graph and the projection process makes the quantification problem theoretically tractable, it is still a challenging issue in practice; and (ii) assuming G^a and G^u are projected from an underlying network implies G^a and G^u have a strong structural correlation. Intuitively, this assumption is reasonable since they correspond to the same group of users and the empirical results in [2][3] also supports such strong structural correlation.

Based on the above assumptions, we have $n!$ possible DA schemes $\sigma : V^a \to V^u$ to de-anonymize G^a, among which the only one *perfect DA scheme* ($\forall i \in V^a$, i is successfully de-anonymized) is denoted by σ_0.

4.2 Model and Formalization

Now, given G, we denote $|V| = n$ and $|E| = m$. Let $V = \{1, 2, \cdots, n\}$ and d_i be the degree of $i \in V$. Then, we define $\mathbf{D} =< d_1, d_2, \cdots, d_n >$ as the degree sequence of the nodes (users) in V. Furthermore, let Δ_1 and Δ_2 (resp., δ_1 and δ_2) be the *maximum* and *second maximum* (resp., *minimum* and *second minimum*) degrees of G, respectively. In [5], Pedarsani and Grossglauser quantified the privacy of G when G is an ER random graph $G(n,p)$ [2]. The $G(n,p)$ model is very useful as a source of insight into the study of structural data, e.g., SNs [5][21]. However, the degree distribution of $G(n,p)$ tends to follow the Poisson distribution, which is quite different from the degree distributions of most, if not all, observed real world structural data (e.g., SNs) [21][22]. Actually, the degree distribution of real world structural data may follow any distribution such as the power-law distribution, exponential distribution, etc. [21][22]. Therefore, it is significant to understand and quantify a DA attack for structural data under an *arbitrary degree distribution*. To this end, we characterize G by a generalized graph model, the *configuration model* [21]. Under the configuration model, a graph is specified by an arbitrary degree sequence $\mathbf{D}$ rather than a particular degree distribution. Since $\mathbf{D}$ is an arbitrary degree sequence, $\mathbf{D}$ can follow an arbitrary distribution observed in real world data [21].

Let $p_{i,j}$ be the probability of an edge existing between $i, j \in V$. Then, we have $p_{i,j} = \frac{d_i d_j}{2m-1} \underset{\text{as } m\to\infty}{\simeq} \frac{d_i d_j}{2m}$, which is a key property of the configuration model [21]. From $p_{i,j}$, it is more likely of an existing edge between two users with high degrees. Based on $p_{i,j}$, we define $l = \min\{p_{i,j} | i, j \in V, i \neq j\}$ and $h = \max\{p_{i,j} | i, j \in V, i \neq j\}$, i.e., l and h are the lower and upper bounds of $p_{i,j}$ respectively. Then, given G with arbitrary degree distribution, we have $l \geq \frac{\delta_1\delta_2}{2m-1}$ and $h \leq \frac{\Delta_1\Delta_2}{2m-1}$.

Finally, given any DA scheme $\sigma = \{(i,i') | 1 \leq i, i' \leq n, i \in V^a, i' \in V^u\} \subseteq V^a \times V^u$, we define the *DA Error* (DE) on a user mapping $(i,i') \in \sigma$ as $\psi_{i,i'} = |N^a_i \setminus N^u_{i'}| + |N^u_{i'} \setminus N^a_i|$, which measures the neighborhoods' difference between i in G^a and i' in G^u under the particular σ. Then, we define the overall DE for a particular σ as $\Psi_\sigma = \sum_{(i,i')\in\sigma} \psi_{i,i'}$. Taking G^a and G^u shown in Fig. 1 as an example, the DE of the perfect DA scheme σ_0 is $\Psi_{\sigma_0} = 20$. For another DA scheme $\sigma = (\sigma_0 \setminus \{(4,4),(5,5)\}) \cup \{(4,5),(5,4)\}$ (users 4 and 5 are incorrectly de-anonymized to each other), its DE is $\Psi_\sigma = 28$. In the following subsections, *we quantify a DA attack by studying the conditions on G and the projection process under which perfect and $(1-\epsilon)$-perfect DA attacks can be conducted.*

4.3 Perfect DA Quantification

Now, we quantify the conditions for perfect DA attacks. Some useful properties of the *binomial distribution* that will be used in the proofs are as follows.

LEMMA 1. *(i) Let $X \sim \mathbf{B}(n_1, p)$ and $Y \sim \mathbf{B}(n_2, p)$ be independent binomial variables. Then, $X + Y$ is again a binomial variable and $X + Y \sim \mathbf{B}(n_1 + n_2, p)$; (ii) [5] Let X and Y be two binomial random variables with means λ_x and*

[2]Based on the projection process, G^a and G^u are also ER random graphs $G(n, p \cdot p_a)$ and $G(n, p \cdot p_u)$, respectively.

λ_y, *respectively. Then, when* $\lambda_x > \lambda_y$, $\Pr(X - Y \leq 0) \leq 2\exp(-\frac{(\lambda_x-\lambda_y)^2}{8(\lambda_x+\lambda_y)})$.

4.3.1 Same Projection Probability

First, we consider the scenario that the projection processes from G to G^a and G^u are characterized by the same probability $\wp$, i.e., $p_a = p_u = \wp$. Let $f_\wp = \frac{\wp[l(1-h\wp)-h(1-\wp)]^2}{2(l(1-h\wp)+h(1-\wp))}$ be a variable depending on $\wp$. Then, we have the following Theorem 1 which indicates the conditions on $\wp$ and $f_\wp$ such that it is *asymptotically almost surely* (*a.a.s.*)[3] that $\Psi_\sigma \geq \Psi_{\sigma_0}$ for any DA scheme $\sigma \neq \sigma_0$. We defer the proof to Appendix A for readability.

THEOREM 1. *For any* $\sigma \neq \sigma_0$, *let* k *be the number of different mappings between* σ *and* σ_0, *i.e., the number of incorrect mappings in* σ. *Then,* $2 \leq k \leq n$ *and* $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \underset{n\to\infty}{\to} 1$ *when* $\wp > \frac{h-l}{h-hl}$ *and* $f_\wp = \Omega(\frac{2\ln n+1}{kn})$.

In Theorem 1, we quantified the condition on $\wp, l$, and h under which the perfect DA scheme σ_0 will cause less DE than any other given DA scheme $\sigma \neq \sigma_0$. To guarantee the *uniqueness* of σ_0 (i.e., σ_0 is *the one and the only one* DA scheme introducing the least DE), intuitively, stronger conditions on $\wp, l$, and h are required. We quantify such conditions in Theorem 2. We defer the proof to Appendix B for readability.

THEOREM 2. *Let* $\mathbf{E}$ *be the event that* there exists at least one DA scheme $\sigma \neq \sigma_0$ such that $\Psi_\sigma \leq \Psi_{\sigma_0}$. *When* $\wp > \frac{h-l}{h-hl}$ *and* $f_\wp = \Omega(\frac{(k+3)\ln n+1}{kn})$, *where* $2 \leq k \leq n$, $\Pr(\mathbf{E}) \to 0$, *i.e., it is* a.a.s. *that* $\nexists\sigma$ *s.t.* $\sigma \neq \sigma_0$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$.

From Theorem 2, although we seek a stronger result, the condition on $\wp$ is the same as in Theorem 1 and the condition on $f_\wp$ only has an increase of order $\Theta(k)$. Based on Theorem 2, if $\wp > \frac{h-l}{h-hl}$ and $f_\wp = \Omega(\frac{(k+3)\ln n+1}{kn})$, the perfect DA scheme causes the least DE. Furthermore, the number of possible DA schemes is upper-bounded. Therefore, when the conditions on $\wp$ and $f_\wp$ are satisfied, G^a can mathematically be perfectly de-anonymized by G^u based on the structure information only.

4.3.2 Different Projection Probabilities

Now, we quantify the conditions on p_a, p_u, l, and h when $p_a \neq p_u$ for structure based perfect DA attacks. Let $g_{p_a,p_u} = \frac{p_a p_u}{p_a+p_u}$ and $f_{p_a,p_u} = \frac{(l(p_a+p_u-2hp_ap_u)-h(p_a+p_u-2p_ap_u))^2}{4(l(p_a+p_u-2hp_ap_u)+h(p_a+p_u-2p_ap_u))}$ be two variables depending on p_a and p_u. Then, we have the following theorems quantifying the conditions on $g_{p_a,p_u}, f_{p_a,p_u}, l$, and h under which it is *a.a.s.* $\Psi_\sigma \geq \Psi_{\sigma_0}$ for any $\sigma \neq \sigma_0$. We omit the proofs due to space limitation.

THEOREM 3. $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \to 1$ *for any* $\sigma \neq \sigma_0$ *when* $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$ *and* $f_{p_a,p_u} = \Omega(\frac{2\ln n+1}{kn})$.

THEOREM 4. *It is* a.a.s. *that* $\nexists\sigma$ *s.t.* $\sigma \neq \sigma_0$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$ *when* $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$ *and* $f_{p_a,p_u} = \Omega(\frac{(k+3)\ln n+1}{kn})$, *where* $2 \leq k \leq n$.

From Theorem 4, to guarantee the uniqueness of inducing the least DE of σ_0, which is a stronger conclusion compared with that in Theorem 3, the condition on g_{p_a,p_u} is the same as in Theorem 3 and the condition on f_{p_a,p_u} has an increase of $\Theta(k)$. Furthermore, Theorem 4 quantifies the conditions under which the anonymized structural data can be mathematically perfectly de-anonymized when $p_a \neq p_u$.

4.4 $(1-\epsilon)$-Perfect DA Quantification

Formally, we define a $(1-\epsilon)$-*perfect DA*, denoted by σ^ϵ, as a DA scheme under which at most $\epsilon|V^a| = \epsilon n$ users are tolerated to be incorrectly de-anonymized, where $0 \leq \epsilon \leq 1$. Under the $(1-\epsilon)$-perfect DA assumption, any σ_k is proper as long as $k \leq \epsilon n$, i.e., we take it as a satisfiable de-anonymizatoin solution. Theoretically, the conditions on $(1-\epsilon)$-perfect DA are quantified in Theorems 5 and 6. Again, we omit the proofs due to space limitation. Note that, when we quantify the conditions for $(1-\epsilon)$-perfect de-aonymization, we do not distinguish σ_0 and σ_k with $k \leq \epsilon n$, since they are all proper solutions. Hence, as in the scenario of perfect DA, our quantification takes σ_0 as the reference point.

THEOREM 5. *(i) When* $p_a = p_u = \wp$, $\wp > \frac{h-l}{h-hl}$, *and* $f_\wp = \Omega(\frac{2\ln n+1}{\epsilon n^2})$, $\Pr(\Psi_{\sigma_k} \geq \Psi_{\sigma_0})$ *for any* σ_k *with* $k > \epsilon n$; *(ii) When* $p_a \neq p_u$, $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$, *and* $f_{p_a,p_u} = \Omega(\frac{2\ln n+1}{\epsilon n^2})$, $\Pr(\Psi_{\sigma_k} \geq \Psi_{\sigma_0})$ *for any* σ_k *with* $k > \epsilon n$.

THEOREM 6. *(i) When* $p_a = p_u = \wp$, $\wp > \frac{h-l}{h-hl}$, *and* $f_\wp = \Omega(\frac{(\epsilon n+3)\ln n+1}{\epsilon n^2})$, *it is* a.a.s. *that there exists no* σ_k *such that* $k > \epsilon n$ *and* $\Psi_{\sigma_k} \leq \Psi_{\sigma_0}$; *(ii) When* $p_a \neq p_u$, $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$, *and* $f_{p_a,p_u} = \Omega(\frac{(\epsilon n+3)\ln n+1}{\epsilon n^2})$, *it is* a.a.s. *that* $\nexists\sigma_k$ *s.t.* $k > \epsilon n$ *and* $\Psi_{\sigma_k} \leq \Psi_{\sigma_0}$.

From Theorem 5, we can see that (i) for any DA scheme σ_k, if it has more than ϵn incorrect mappings, with probability 1, it will cause more DE than σ_0. On the other hand, if σ_k is a $(1-\epsilon)$-perfect DA scheme, i.e., $k \leq \epsilon n$, we cannot *a.a.s.* distinguish σ_k and σ_0 based on DE under the quantified conditions; (ii) compared with the quantifications in Theorems 1 and 3, the conditions on $f_\wp$ and f_{p_a,p_u} change from $\Omega(\frac{\ln n}{kn})$ to $\Omega(\frac{\ln n}{n^2})$ explicitly, which implies a relaxation of the condition on $f_\wp$ and f_{p_a,p_u}. This relaxation comes from the toleration of ϵn incorrect user mappings. As in the scenario of perfect DA, stronger conditions can be quantified to guarantee $(1-\epsilon)$-perfect DA schemes causing the least DE as shown in Theorem 6. From Theorem 6, we can see that even ϵn matching errors are tolerated, the conditions on $\wp$ and g_{p_a,p_u} stay the same while the conditions on $f_\wp$ and f_{p_a,p_u} only have some constant relaxation compared with the perfect DA scenario.

5. LARGE SCALE EVALUATION ON REAL WORLD STRUCTURAL DATASETS

According to our quantification, even without semantic priori knowledge, anonymized structural data can be de-anonymized perfectly or $(1-\epsilon)$-perfectly. In this section, we conduct comprehensive evaluations of our DA quantification on 26 real world structural datasets[4].

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

[4] We actually conduct evaluations on 60+ real world datasets. Due to space limitation, the results on 26 representative datasets are shown in the paper. Complete results and source codes are available up to request.

Table 1: Data statistics.

Name	Type	n	m	ρ	$\overline{d}$	$p(1)$	$p(5)$
Google+	SN	4.7M	90.8M	8.24E-6	38.7	.054	.273
Twitter	SN	.5M	14.9M	1.20E-4	54.8	.053	.198
LiveJournal	SN	4.8M	69M	3.70E-6	17.9	.210	.505
Facebook	SN	4K	88K	1.08E-2	43.7	.019	.113
YouTube	SN	1.1M	3M	4.64E-6	5.3	.531	.855
Orkut	SN	3.1M	117.2M	2.48E-5	76.3	.022	.073
Slashdot	SN	82.2K	1M	1.73E-4	14.2	.022	.593
Pokec	SN	1.6M	30.6M	1.67E-5	27.3	.100	.307
Infocom	LMSN	73	212	8.07E-2	5.8	.068	.493
Smallblue	LMSN	120	375	5.25E-2	6.3	.133	.625
Brightkite	LMSN	58K	.2M	1.32E-4	7.5	.354	.718
Gowalla	LMSN	.2M	1M	4.92E-5	9.7	.252	.645
HepPh	ColN	12K	.2M	1.87E-3	21.0	.100	.500
AstroPh	ColN	18.8K	.4M	1.23E-3	22.0	.053	.337
CondMat	ColN	23.1K	.2M	4.00E-4	8.6	.078	.518
DBLP	ColN	.3M	1.1M	2.09E-5	6.6	.136	.670
Enron	Email	36.7K	.2M	3.19E-4	10.7	.281	.679
EuAll	Email	.3M	.4M	1.35E-5	3.0	.837	.973
Wiki	WikiTalk	2.4M	5M	1.63E-6	3.9	.738	.962
AS733	AS	6.5K	13.9K	6.63E-4	4.3	.355	.896
Oregon	AS	11.5K	32.7K	4.98E-4	5.7	.289	.876
Caida	AS	26.5K	53.4K	1.52E-4	4.0	.375	.924
Skitter	AS	1.7M	11.1M	7.73E-6	13.1	.128	.554
Gnutella3	P2P	26.5K	65.4K	1.86E-4	4.9	.413	.710
Gnutella4	P2P	36.7K	88.3K	1.32E-4	4.8	.448	.718
Gnutella5	P2P	62.6K	.1M	7.56E-5	4.7	.458	.725

5.1 Evaluation Setup

During the quantification, $p_{i,j}$ is an important parameter although we quantify the conditions in laconic expressions in terms of its bounds l and h. However, it is difficult to accurately determine $p_{i,j}$ in practice [5][21][27]. Fortunately, it is not necessary to know the exact $p_{i,j}$ to numerically evaluate our DA quantification. Actually, according to our derivation, we only have to determine the statistical *expectation value* of $p_{i,j}$, denoted by $\mathbb{E}(p_{i,j})$. For a dataset with degree sequence $\mathbf{D}$, define $p_{\mathbf{D}} = \mathbb{E}(p_{i,j})$. Then, it is statistically reasonable (especially for large datasets) to use the *graph density* $\rho = \frac{2m}{n(n-1)}$ to approximate $p_{\mathbf{D}}$, i.e., $p_{\mathbf{D}} \simeq \rho$ [5][21]. On the other hand, we focus on demonstrating the statistical behavior of our perfect/$(1-\epsilon)$-perfect DA quantification. Therefore, we use ρ to approximate $p_{\mathbf{D}}$ in our evaluation. Furthermore, for the convenience of evaluation, we evaluate the quantification in the scenario of $p_a = p_u = \wp$. This does not limit our evaluation since it is straightforward to extend to the $p_a \neq p_u$ scenario (actually, both scenarios exhibit similar behaviors, which can also be seen in the quantification).

Let $f_{\mathbf{D}} = \frac{p_{\mathbf{D}}\wp(\wp - p_{\mathbf{D}}\wp)^2}{2(2-p_{\mathbf{D}}\wp-\wp)}$. Then, we have the following conclusions, which can be proven by similar techniques as in Theorems 1, 2, 5, and 6 from the statistical perspective.

THEOREM 7. *For perfect DA, (i)* $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \to 1$ *for any* $\sigma \neq \sigma_0$ *when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{4\ln n+2}{2kn-k^2-k})$*; (ii) it is* a.a.s. *that* $\nexists\sigma$ *s.t.* $\sigma \neq \sigma_0$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$ *when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{2(k+3)\ln n+2}{2kn-k^2-k})$.

THEOREM 8. *For* $(1-\epsilon)$*-perfect DA, (i)* $\Pr(\Psi_{\sigma_k} \geq \Psi_{\sigma_0}) \to 1$ *for any* σ_k *with* $k > \epsilon n$ *when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{\ln n}{\epsilon n^2})$*; (ii) it is* a.a.s. *that* $\nexists\sigma_k$ *s.t.* $k > \epsilon n$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$ *when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{\ln n}{n})$.

5.2 Datasets

We evaluate our quantification on 26 datasets from multiple domains, including SNs data, Location based Mobility traces and SN (LMSN) data, Collaboration Network (ColN) data, communication network (Email, WikiTalk) data, Autonomous Systems (AS) graph data, and Peer-to-Peer (P2P) network graph data [3][25][26][27]. In Tab. 1, we show some statistics on the employed datasets, where $\overline{d}$ represents the *average degree* of n nodes and $p(i)$ indicates the *percentage* of nodes with degree of i or less.

Due to space limitation, we briefly introduce the datasets as follows. Detailed descriptions can be found in [3][25][26][27]. **SN:** Google+, Twitter, LiveJournal, Facebook, YouTube, Orkut, Slashdot, and Pokec are 8 well known SNs [25][26]. **LMSN:** Infocom consists of a Bluetooth contact trace and Smallblue consists of an *instant messenger* contact trace [3]. Both Brightkite and Gowalla consist of a SN and a check-in trace of the SN users [26][27]. **ColN:** HepPh, AstroPh, and CondMat are three collaboration networks from arXiv in the areas of *High Energy Physics-Phenomenology*, *Astro Physics*, and *Condense Matter Physics*, respectively [26]. DBLP is a collaboration network of researchers mainly in *Computer Science* [26]. **Email and WikiTalk:** Enron and EuAll are two email communication networks [26]. WikiTalk is a network containing the discussion relationships among a group of users on Wikipedia [26]. **AS:** AS733, Oregon, Caida, and Skitter are four AS graphs at different locations [26]. **P2P:** Gnutella3, Gnutella4, and Gnutella5 are three P2P network graphs where nodes represent hosts in Gnutella and edges are connections between hosts [26].

5.3 Evaluation on Perfect DA Quantification

The conditions in Theorems 7 and 8 are quantified in sense of n being a large number. Therefore, in the evaluation of perfect/$(1-\epsilon)$-perfect DA quantification, we derive an extra condition on the lower bound on n, denoted by $\Omega(n)$. Then, based on Theorem 7, the conditions on $(\Omega(f_{\mathbf{D}}), \Omega(n))$ for perfect DA under different $\wp$ are shown in Tab. 2. From Tab. 2, we have the following observations.

(*i*) When $\wp$ increases, $\Omega(f_{\mathbf{D}})$ shows an increasing trend. For instance, $\Omega(f_{\mathbf{D}})$ is increased from 6.5E-8 when $\wp = .3$ to 2.7E-6 when $\wp = .9$, which implies the condition on $f_{\mathbf{D}}$ becomes stronger. This is consistent with our quantification since $f_{\mathbf{D}}$ is an *increasing function* on $\wp$ given $p_{\mathbf{D}}$. On the other hand, we find that although $\Omega(f_{\mathbf{D}})$ increases for large $\wp$, it still keeps relatively loose bounds, i.e., $f_{\mathbf{D}}$ is easily satisfied. For example, when $\wp = .9$, the condition on $\Omega(f_{\mathbf{D}})$ is 2.7E-6 for Google+ (a large scale dataset) and 1.6E-5 for Gowalla (a medium scale dataset).

(*ii*) When $\wp$ increases, $\Omega(n)$ decreases. For instance, $\Omega(n)$ is decreased from 1.7E7 when $\wp = .3$ to 3.2E5 when $\wp = .9$ for Twitter. This is because a large $\wp$ implies that G^a is topologically more similar to G^u. Thus, a weaker condition on $\Omega(n)$ is sufficient to enable a perfect DA scheme *a.a.s.* inducing the least DE.

(*iii*) For datasets with similar graph densities, e.g., Google+ (ρ = 8.24E-6) and Skitter (ρ = 7.73E-6), the conditions on $(\Omega(f_{\mathbf{D}}), \Omega(n))$ are also similar for perfect DA, which is consistent with our theoretical quantification. This comes from the similarity of their statistical $p_{\mathbf{D}}$. For perfect DA on datasets with different graph densities (with similar or different sizes), e.g., HepPh (n = 1.2E4, ρ = 1.87E-3) and

Table 2: Evaluation of $(\Omega(f_{\mathbf{D}}), \Omega(n))$ in perfect DA.

Dataset	n	$\wp = .3$	$\wp = .4$	$\wp = .5$	$\wp = .6$	$\wp = .7$	$\wp = .8$	$\wp = .9$
Google+	4.7E6	(6.5E-8, 3.0E8)	(1.6E-7, 1.1E8)	(3.4E-7, 5.2E7)	(6.4E-7, 2.7E7)	(1.1E-6, 1.5E7)	(1.8E-6, 9.1E6)	(2.7E-6, 5.7E6)
Twitter	4.6E5	(9.5E-7, 1.7E7)	(2.4E-6, 6.5E6)	(5.0E-6, 3.0E6)	(9.3E-6, 1.5E6)	(1.6E-5, 8.6E5)	(2.6E-5, 5.1E5)	(4.0E-5, 3.2E5)
LiveJournal	4.8E6	(2.9E-8, 6.9E8)	(7.4E-8, 2.6E8)	(1.5E-7, 1.2E8)	(2.9E-7, 6.3E7)	(4.9E-7, 3.6E7)	(7.9E-7, 2.1E7)	(1.2E-6, 1.3E7)
Facebook	4.0E3	(8.4E-5, 1.4E5)	(2.1E-4, 5.1E4)	(4.4E-4, 2.3E4)	(8.2E-4, 1.1E4)	(1.4E-3, 6.2E3)	(2.3E-3, 3.6E3)	(3.5E-3, 2.2E3)
YouTube	1.1E6	(3.7E-8, 5.5E8)	(9.3E-8, 2.1E8)	(1.9E-7, 9.5E7)	(3.6E-7, 5.0E7)	(6.1E-7, 2.8E7)	(9.9E-7, 1.7E7)	(1.5E-6, 1.1E7)
Orkut	3.1E6	(2.0E-7, 9.3E7)	(5.0E-7, 3.5E7)	(1.0E-6, 1.6E7)	(1.9E-6, 8.3E6)	(3.3E-6, 4.7E6)	(5.3E-6, 2.8E6)	(8.2E-6, 1.7E6)
Slashdot	8.2E4	(1.4E-6, 1.2E7)	(3.5E-6, 4.4E6)	(7.2E-6, 2.0E6)	(1.3E-5, 1.0E6)	(2.3E-5, 5.8E5)	(3.7E-5, 3.5E5)	(5.7E-5, 2.1E5)
Pokec	1.6E6	(1.3E-7, 1.4E8)	(3.3E-7, 5.3E7)	(7.0E-7, 2.4E7)	(1.3E-6, 1.3E7)	(2.2E-6, 7.2E6)	(3.6E-6, 4.3E6)	(5.5E-6, 2.7E6)
Infocom	7.3E1	(5.5E-4, 1.8E4)	(1.4E-3, 6.4E3)	(2.9E-3, 2.7E3)	(5.4E-3, 1.4E3)	(9.4E-3, 7.8E2)	(1.5E-2, 3.9E2)	(2.4E-2, 2.5E2)
Smallblue	1.2E2	(3.8E-4, 2.7E4)	(9.6E-4, 9.7E3)	(2.0E-3, 4.2E3)	(3.7E-3, 2.1E3)	(6.4E-3, 1.2E3)	(1.0E-2, 6.8E2)	(1.6E-2, 4.4E2)
Brightkite	5.7E4	(1.1E-6, 1.6E7)	(2.6E-6, 5.9E6)	(5.5E-6, 2.7E6)	(1.0E-5, 1.4E6)	(1.7E-5, 7.8E5)	(2.8E-5, 4.6E5)	(4.4E-5, 2.9E5)
Gowalla	2.0E5	(3.9E-7, 4.5E7)	(9.8E-7, 1.7E7)	(2.0E-6, 7.7E6)	(3.8E-6, 4.0E6)	(6.5E-6, 2.3E6)	(1.0E-5, 1.3E6)	(1.6E-5, 8.4E5)
HepPh	1.2E4	(1.5E-5, 9.3E5)	(3.7E-5, 3.4E5)	(7.8E-5, 1.5E5)	(1.4E-4, 7.8E4)	(2.5E-4, 4.3E4)	(4.0E-4, 2.6E4)	(6.2E-4, 1.6E4)
AstroPh	1.8E4	(9.7E-6, 1.5E6)	(2.5E-5, 5.4E5)	(5.1E-5, 2.4E5)	(9.5E-5, 1.2E5)	(1.6E-4, 6.9E4)	(2.6E-4, 4.1E4)	(4.1E-4, 2.5E4)
CondMat	2.1E4	(3.2E-6, 4.8E6)	(8.0E-6, 1.8E6)	(1.7E-5, 8.2E5)	(3.1E-5, 4.2E5)	(5.3E-5, 2.3E5)	(8.5E-5, 1.4E5)	(1.3E-4, 8.6E4)
DBLP	3.2E5	(1.7E-7, 1.1E8)	(4.2E-7, 4.2E7)	(8.7E-7, 1.9E7)	(1.6E-6, 1.0E7)	(2.8E-6, 5.6E6)	(4.5E-6, 3.4E6)	(6.9E-6, 2.1E6)
Enron	3.4E4	(2.5E-6, 6.2E6)	(6.4E-6, 2.3E6)	(1.3E-5, 1.0E6)	(2.5E-5, 5.4E5)	(4.2E-5, 3.0E5)	(6.8E-5, 1.8E5)	(1.1E-4, 1.1E5)
EuAll	2.2E5	(1.1E-7, 1.8E8)	(2.7E-7, 6.7E7)	(5.6E-7, 3.1E7)	(1.0E-6, 1.6E7)	(1.8E-6, 9.0E6)	(2.9E-6, 5.4E6)	(4.5E-6, 3.4E6)
Wiki	2.4E6	(1.3E-8, 1.6E9)	(3.3E-8, 6.2E8)	(6.8E-8, 2.9E8)	(1.3E-7, 1.5E8)	(2.2E-7, 8.5E7)	(3.5E-7, 5.1E7)	(5.4E-7, 3.2E7)
AS733	6.5E3	(5.3E-6, 2.8E6)	(1.3E-5, 1.0E6)	(2.8E-5, 4.7E5)	(5.1E-5, 2.4E5)	(8.7E-5, 1.4E5)	(1.4E-4, 8.0E4)	(2.2E-4, 4.9E4)
Oregon	1.1E4	(4.0E-6, 3.8E6)	(1.0E-5, 1.4E6)	(2.1E-5, 6.4E5)	(3.8E-5, 3.3E5)	(6.6E-5, 1.8E5)	(1.1E-4, 1.1E5)	(1.7E-4, 6.7E4)
Caida	2.6E4	(1.2E-6, 1.4E7)	(3.0E-6, 5.1E6)	(6.3E-6, 2.3E6)	(1.2E-5, 1.2E6)	(2.0E-5, 6.7E5)	(3.2E-5, 4.0E5)	(5.0E-5, 2.5E5)
Skitter	1.7E6	(6.1E-8, 3.2E8)	(1.5E-7, 1.2E8)	(3.2E-7, 5.5E7)	(6.0E-7, 2.9E7)	(1.0E-6, 1.6E7)	(1.6E-6, 9.8E6)	(2.6E-6, 6.1E6)
Gnutella3	2.6E4	(1.5E-6, 1.1E7)	(3.7E-6, 4.1E6)	(7.8E-6, 1.9E6)	(1.4E-5, 9.6E5)	(2.5E-5, 5.4E5)	(4.0E-5, 3.2E5)	(6.2E-5, 2.0E5)
Gnutella4	3.7E4	(1.0E-6, 1.6E7)	(2.6E-6, 5.9E6)	(5.5E-6, 2.7E6)	(1.0E-5, 1.4E6)	(1.7E-5, 7.8E5)	(2.8E-5, 4.7E5)	(4.4E-5, 2.9E5)
Gnutella5	6.3E4	(6.0E-7, 2.9E7)	(1.5E-6, 1.1E7)	(3.1E-6, 4.9E6)	(5.8E-6, 2.5E6)	(1.0E-5, 1.4E6)	(1.6E-5, 8.5E5)	(2.5E-5, 5.3E5)

Oregon (n = 1.15E4, ρ = 4.98E-4), Facebook (n = 4.0E3, ρ = 1.08E-2) and Twitter (n = 4.6E5, ρ = 1.2E-4), dense datasets require a stronger condition on $f_{\mathbf{D}}$ while a weaker condition on $\Omega(n)$ given $\wp$, which is also consistent with our quantification. A stronger condition requirement on $f_{\mathbf{D}}$ is because $f_{\mathbf{D}}$ is an increasing function on $p_{\mathcal{D}} \simeq \rho \in (0, 0.5]$ given $\wp$ and all the considering datasets have $\rho \leq 0.5$. A looser bound on $\Omega(n)$ comes from the fact that more structural information can be projected to G^a and G^u in dense datasets.

(*iv*) From Tab. 2, some datasets can be perfectly de-anonymized under some conditions. For instance, Orkut and Facebook are *a.a.s.* can be perfectly de-anonymized when $\wp \geq \Omega(.8)$. The perfect DA is due to their good structural characteristics, e.g., high average degree, small percentage of nodes with a low degree.

5.4 Evaluation on $(1-\epsilon)$-Perfect DA Quantification

Based on our quantification, the percentage of successfully de-anonymized users by any $(1-\epsilon)$-perfect DA scheme is at least $1-\epsilon$. Given $\wp$ varied from .3 to .95, we evaluate the minimum number of users in the 26 datasets considered that can be successfully de-anonymized with probability 1 in terms of our quantification, i.e., the lower bound of $1-\epsilon$, $(\Omega(1-\epsilon))$, and the results are shown in Tab. 3. From Tab. 3, we make some important observations and comments as follows.

(*i*) When $\wp$ increases, more users can be de-anonymized for every dataset as expected. For example, when $\wp$ = .5, it is *a.a.s.* at least 29.7% of the users in Google+ can be successfully de-aonymized; when $\wp$ is increased to .8, at least 72.5% of the users in Google+ can be successfully de-anonymized; when $\wp$ = .95 all the users in Google+ can *a.a.s.* be successfully de-anonymized. From Tab. 3, similar DA phenomena applied to all the datasets, which is consistent with our quantification. The reason is straightforward. When $\wp$ increases, more edges appear in both G^a and G^u. Thus, the structural similarity between G^a and G^u is increased followed by more users can statistically be successfully de-anonymized.

(*ii*) Most of the existing structural datasets are *a.a.s.* de-anonymizable completely or at least partially just based on the topological information. For instance, Facebook and Orkut datasets can be completely de-anonymized when $\wp$ = .8. Even when a dataset cannot be completely de-anonymized, it may be de-anonymizable partially in a large-scale. For example, when $\wp$ = .9, at least 60.9%, 48.9%, and 85.7% of the users in LiveJournal, Gowalla, and AstroPh can be successfully de-anonymized, respectively. This fact is consistent with our quantification as well as the intuition that structure itself can be used to de-anonymize data.

(*iii*) An interesting observation is that the DA results on two datasets with similar graph densities may be very different in practice. From Tab. 2, for two datasets with similar graph densities, e.g., Google+ and Skitter, the theoretical bounds on $(\Omega(f_{\mathbf{D}}), \Omega(n))$ for perfect DA are also similar. However, from Tab. 3, the DA results of Google+ and Skitter are very different: when $\wp$ = .6, the number of de-anonymizable users in Google+ (41.8%) is about twice that of in Skitter (23.1%); while when $\wp$ = .95, all the users in Google+ are *a.a.s.* de-anonymizable while the de-anonymizable users in Skitter is only bounded by $\Omega(59.1\%)$. To study the reason for this fact, we need to consider the degree distribution of Google+ and Skitter besides the graph density (as well as $\Omega(f_{\mathbf{D}})$ and $\Omega(n)$). From Tab. 1, the percentage of low degree users in Skitter is much higher than that in Google+. On the other hand, intuitively, low degree users, especially users with degree of 1, do not have too much distinguishable structural information (this intuition is confirmed by our theoretical quantification on different DEs caused by mismatching high degree users and low degree users), which implies that they are difficult to be de-anonymized based on structural information. Consequently,

Table 3: Evaluation of $\Omega(1-\epsilon)$ in $(1-\epsilon)$-perfect DA.

Dataset	$\wp = .3$	$\wp = .35$	$\wp = .4$	$\wp = .45$	$\wp = .5$	$\wp = .55$	$\wp = .6$	$\wp = .65$	$\wp = .7$	$\wp = .75$	$\wp = .8$	$\wp = .85$	$\wp = .9$	$\wp = .95$
Google+	11.7%	15.5%	19.7%	24.5%	29.7%	35.5%	41.8%	48.7%	56.1%	64.0%	72.5%	81.6%	91.2%	100.0%
Twitter	15.1%	20.2%	26.0%	32.4%	39.4%	47.1%	55.4%	64.3%	73.8%	84.0%	94.7%	100.0%	100.0%	100.0%
LiveJournal	6.6%	9.1%	11.9%	15.2%	18.8%	22.7%	27.1%	31.8%	36.8%	42.3%	48.1%	54.3%	60.9%	68.1%
Facebook	3.7%	12.1%	22.4%	31.0%	39.9%	49.5%	59.6%	70.3%	81.5%	93.2%	100.0%	100.0%	100.0%	100.0%
YouTube	4.0%	5.3%	6.8%	8.4%	10.3%	12.3%	14.5%	16.9%	19.5%	22.4%	25.5%	28.9%	32.5%	36.4%
Orkut	14.2%	19.6%	26.0%	33.3%	41.4%	50.3%	60.0%	70.4%	81.3%	92.7%	100.0%	100.0%	100.0%	100.0%
Slashdot	7.2%	9.8%	12.7%	15.9%	19.5%	23.4%	27.6%	32.2%	37.2%	42.7%	48.6%	54.9%	61.8%	69.3%
Pokec	7.3%	10.4%	14.1%	18.4%	23.2%	28.5%	34.4%	40.7%	47.5%	54.7%	62.4%	70.5%	79.0%	88.1%
Infocom	10.4%	11.5%	12.5%	13.0%	13.9%	14.3%	15.1%	15.5%	15.8%	16.6%	16.9%	17.2%	49.9%	62.2%
Smallblue	8.9%	9.6%	10.3%	10.9%	11.3%	11.8%	12.1%	12.6%	12.9%	13.3%	33.0%	44.6%	54.6%	64.7%
Brightkite	4.7%	6.5%	8.6%	10.9%	13.5%	16.4%	19.6%	23.1%	26.8%	30.9%	35.3%	40.0%	45.1%	50.6%
Gowalla	5.3%	7.2%	9.4%	11.9%	14.7%	17.8%	21.2%	25.0%	29.0%	33.4%	38.2%	43.3%	48.9%	54.8%
HepPh	9.0%	13.2%	17.6%	22.4%	27.6%	33.2%	39.2%	45.7%	52.7%	60.1%	68.1%	76.7%	85.9%	95.7%
AstroPh	7.4%	11.0%	15.3%	20.1%	25.4%	31.2%	37.6%	44.4%	51.7%	59.4%	67.6%	76.4%	85.7%	95.6%
CondMat	3.5%	5.2%	7.2%	9.6%	12.3%	15.3%	18.7%	22.6%	26.8%	31.4%	36.5%	42.1%	48.2%	54.8%
DBLP	3.0%	4.3%	5.8%	7.6%	9.6%	11.8%	14.3%	17.1%	20.2%	23.6%	27.4%	31.5%	36.0%	40.9%
Enron	6.6%	9.0%	11.7%	14.6%	17.9%	21.4%	25.3%	29.5%	34.1%	39.1%	44.5%	50.3%	56.6%	63.4%
EuAll	3.5%	4.5%	5.6%	6.9%	8.3%	9.8%	11.4%	13.3%	15.2%	17.4%	19.6%	22.1%	24.7%	27.6%
Wiki	3.7%	4.8%	6.0%	7.4%	8.9%	10.5%	12.3%	14.2%	16.3%	18.6%	21.1%	23.8%	26.7%	29.8%
AS733	1.3%	4.8%	6.5%	8.3%	10.3%	12.5%	14.9%	17.6%	20.5%	23.8%	27.4%	31.2%	35.5%	40.0%
Oregon	4.6%	6.5%	8.6%	10.8%	13.1%	15.7%	18.5%	21.6%	24.9%	28.6%	32.5%	36.7%	41.3%	46.3%
Caida	3.8%	5.1%	6.5%	8.1%	9.9%	11.8%	14.0%	16.3%	18.8%	21.6%	24.6%	27.8%	31.4%	35.3%
Skitter	6.2%	8.3%	10.6%	13.3%	16.2%	19.5%	23.1%	27.1%	31.4%	36.1%	41.2%	46.7%	52.6%	59.1%
Gnutella3	1.7%	2.6%	3.8%	5.4%	7.2%	9.5%	12.1%	15.2%	18.8%	23.0%	27.3%	31.5%	36.0%	40.6%
Gnutella4	1.8%	2.8%	4.0%	5.5%	7.3%	9.4%	12.0%	15.0%	18.4%	22.5%	26.7%	30.8%	35.1%	39.6%
Gnutella5	1.8%	2.7%	3.9%	5.3%	7.0%	9.1%	11.5%	14.4%	17.7%	21.6%	25.7%	29.7%	33.8%	38.1%

the existence of a large amount of low degree users in Skitter makes it less de-anonymizable than Google+, which is consistent with our quantification. In summary, from Tabs. 1 and 3, if a dataset has a high average degree and a small percentage of low degree users, it is easier to de-anonymize and a large amount of its users are *a.a.s.* de-anonymizable; otherwise, for datasets with a low average degree and a large percentage of low degree users, they are difficult to de-anonymize based solely on structural information.

(*iv*) Following the above observation, we find that there exists some differences between theory and practice on the dominating factor of DA. Theoretically, the graph density plays a dominating factor on determining the bound of $(\Omega(f_{\mathbf{D}}), \Omega(n))$. In practice, the degree distribution and the average degree have more impact on the DA results. This is mainly because we study the quantification from an asymptotical sense in the theoretical scenario (i.e., $n \to \infty$) and the key parameter $p_{i,j}$ asymptotically converges to graph density ρ, i.e., $\mathbb{E}(p_{i,j}) \underset{n \to \infty}{\simeq} \rho$. On the other hand, when quantifying the percentage of de-anonymizable users for each dataset, the actual degree sequence/distribution $\mathbf{D}$ is used to examine when the DA conditions are satisfied.

We also evaluate the impact of $\wp$ and ϵ on the bound of $\Omega(n)$ in $(1-\epsilon)$-perfect DA (we do not show $\Omega(f_{\mathbf{D}})$ since it depends on $\wp$ and exhibits the same behavior as in the perfect DA) as shown in Tab. 4. From Tab. 4, we have the following observations.

(*i*) When ϵ is fixed, the impact of $\wp$ on $\Omega(n)$ in $(1-\epsilon)$-perfect DA is similar to that in perfect DA, i.e., when $\wp$ increases, $\Omega(n)$ decreases. The reason is also the same as before since a large $\wp$ implies more similarity between G^a and G^u and thus a loose condition on $\Omega(n)$ is sufficient to enable σ_k $(k \leq \epsilon n)$ to induce less DE than $\sigma_{k'}$ $(k' > \epsilon n)$.

(*ii*) When $\wp$ is fixed, $\Omega(n)$ is also decreasing with the increase of ϵ. For instance, when $\wp = 0.6$, $\Omega(n)$ decreases from 2.2E7 to 9.5E6 for Google+ when ϵ increases from .1 to .4. This is because of that when ϵ increases, more DE is tolerated, and thus loose condition is required for $\Omega(n)$ to distinguish σ_k $(k \leq \epsilon n)$ and $\sigma_{k'}$ $(k' > \epsilon n)$, which is consistent with our quantification.

(*iii*) As in the perfect DA scenario, graph density is an important factor to impact $\Omega(n)$. Datasets with similar graph density, e.g., Google+ and Skitter, exhibits similar requirement on $\Omega(n)$. A dataset with high graph density, e.g., Facebook and HepPh, corresponds to a loose bound on $\Omega(n)$. The reason is also the same as before.

Table 4: Evaluation of $\Omega(n)$ in $(1-\epsilon)$-perfect DA.

Dataset	$\wp = .6$				$\wp = .9$			
	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$
Google+	2.2E7	1.7E7	1.3E7	9.5E6	4.6E6	3.6E6	2.7E6	2.3E6
Twitter	1.2E6	9.6E5	7.3E5	5.4E5	2.5E5	2.3E5	2.3E5	2.3E5
LiveJournal	5.1E7	4.0E7	3.0E7	2.2E7	1.1E7	8.4E6	6.4E6	4.7E6
Facebook	9.2E3	7.2E3	5.5E3	4.1E3	2.0E3	2.0E3	2.0E3	2.0E3
YouTube	4.0E7	3.2E7	2.5E7	1.8E7	8.6E6	6.8E6	5.2E6	3.8E6
Orkut	6.7E6	5.3E6	4.1E6	3.1E6	1.5E6	1.5E6	1.5E6	1.5E6
Slashdot	8.5E5	6.7E5	5.2E5	3.8E5	1.7E5	1.4E5	1.1E5	7.7E4
Pokec	1.0E7	8.0E6	6.1E6	4.5E6	2.1E6	1.7E6	1.3E6	9.4E5
Infocom	1.2E3	9.8E2	7.8E2	6.8E2	2.5E2	2.5E2	1.7E2	1.7E2
Smallblue	1.8E3	1.4E3	1.2E3	8.8E2	3.4E2	3.2E2	2.2E2	2.2E2
Brightkite	1.1E6	8.8E5	6.7E5	4.9E5	2.3E5	1.8E5	1.4E5	1.0E5
Gowalla	3.2E6	2.5E6	1.9E6	1.4E6	6.7E5	5.3E5	4.0E5	3.0E5
HepPh	6.2E4	4.9E4	3.7E4	2.7E4	1.2E4	9.7E3	7.3E3	5.6E3
AstroPh	9.9E4	7.8E4	5.9E4	4.4E4	2.0E4	1.6E4	1.2E4	9.0E3
CondMat	3.4E5	2.7E5	2.1E5	1.6E5	6.9E4	5.5E4	4.2E4	3.2E4
DBLP	8.1E6	6.5E6	5.0E6	3.8E6	1.7E6	1.4E6	1.1E6	8.0E5
Enron	4.3E5	3.4E5	2.6E5	1.9E5	8.8E4	6.9E4	5.2E4	3.8E4
EuAll	1.3E7	1.1E7	8.3E6	6.2E6	2.8E6	2.2E6	1.7E6	1.3E6
Wiki	1.2E8	9.9E7	7.7E7	5.7E7	2.6E7	2.1E7	1.6E7	1.2E7
AS733	2.0E5	1.6E5	1.2E5	9.0E4	4.0E4	3.2E4	2.4E4	1.8E4
Oregon	2.7E5	2.1E5	1.6E5	1.2E5	5.5E4	4.3E4	3.3E4	2.4E4
Caida	9.8E5	7.8E5	6.0E5	4.5E5	2.0E5	1.6E5	1.2E5	9.1E4
Skitter	2.3E7	1.8E7	1.4E7	1.0E7	4.9E6	3.9E6	3.0E6	2.2E6
Gnutella3	7.8E5	6.2E5	4.8E5	3.5E5	1.6E5	1.3E5	9.7E4	7.1E4
Gnutella4	1.1E6	9.0E5	6.9E5	5.1E5	2.3E5	1.9E5	1.4E5	1.0E5
Gnutella5	2.1E6	1.6E6	1.3E6	9.3E5	4.3E5	3.4E5	2.6E5	1.9E5

Table 5: Evaluation of $(\Omega(\wp), \Omega(f_{\mathbf{D}}), \Omega(n))$ in $(1-\epsilon)$-perfect DA.

Dataset	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$	$\epsilon = .5$
Google+	(1.1E-7, 3.5E-6, 2.2E28)	(1.2E-7, 3.7E-6, 1.8E28)	(1.3E-7, 3.9E-6, 1.6E28)	(1.4E-7, 4.2E-6, 1.3E28)	(1.4E-7, 4.4E-6, 1.1E28)
Twitter	(1.2E-6, 3.1E-5, 1.2E24)	(1.2E-6, 3.2E-5, 9.7E23)	(1.3E-6, 3.4E-5, 8.3E23)	(1.4E-6, 3.6E-5, 6.9E23)	(1.5E-6, 3.8E-5, 5.8E23)
LiveJournal	(1.2E-7, 3.6E-6, 4.7E28)	(1.2E-7, 3.6E-6, 4.7E28)	(1.2E-7, 3.8E-6, 3.8E28)	(1.3E-7, 4.1E-6, 3.0E28)	(1.4E-7, 4.3E-6, 2.7E28)
Facebook	(1.3E-4, 2.2E-3, 5.9E15)	(1.4E-4, 2.3E-3, 5.0E15)	(1.5E-4, 2.5E-3, 4.1E15)	(1.6E-4, 2.6E-3, 3.5E15)	(1.7E-4, 2.8E-3, 2.9E15)
YouTube	(6.0E-7, 1.7E-5, 2.4E26)	(6.0E-7, 1.7E-5, 2.4E26)	(6.0E-7, 1.7E-5, 2.4E26)	(6.0E-7, 1.7E-5, 2.4E26)	(6.0E-7, 1.7E-5, 2.4E26)
Orkut	(1.7E-7, 5.1E-6, 2.0E27)	(1.8E-7, 5.4E-6, 1.7E27)	(1.9E-7, 5.7E-6, 1.4E27)	(2.0E-7, 6.1E-6, 1.2E27)	(2.2E-7, 6.5E-6, 9.8E26)
Slashdot	(7.4E-6, 1.7E-4, 2.8E21)	(7.4E-6, 1.7E-4, 2.8E21)	(7.4E-6, 1.7E-4, 2.8E21)	(8.2E-6, 1.9E-4, 2.1E21)	(8.2E-6, 1.9E-4, 2.1E21)
Pokec	(3.2E-7, 9.2E-6, 4.4E26)	(3.5E-7, 9.9E-6, 3.6E26)	(3.6E-7, 1.0E-5, 3.1E26)	(3.9E-7, 1.1E-5, 2.5E26)	(4.1E-7, 1.2E-5, 2.1E26)
Infocom	(8.6E-3, 8.0E-2, 2.0E09)	(8.6E-3, 8.0E-2, 2.0E09)	(9.1E-3, 8.1E-2, 1.7E09)	(1.0E-2, 8.6E-2, 1.2E09)	(1.1E-2, 8.9E-2, 1.1E09)
Smallblue	(4.7E-3, 5.2E-2, 1.9E10)	(5.1E-3, 5.1E-2, 1.5E10)	(5.3E-3, 5.2E-2, 1.3E10)	(5.8E-3, 5.6E-2, 1.0E10)	(6.4E-3, 6.1E-2, 7.2E09)
Brightkite	(1.1E-5, 2.3E-4, 1.2E21)	(1.1E-5, 2.3E-4, 1.2E21)	(1.1E-5, 2.3E-4, 1.2E21)	(1.2E-5, 2.6E-4, 8.7E20)	(1.2E-5, 2.6E-4, 8.7E20)
Gowalla	(2.9E-6, 7.1E-5, 1.8E23)	(2.9E-6, 7.1E-5, 1.8E23)	(3.2E-6, 7.8E-5, 1.3E23)	(3.2E-6, 7.8E-5, 1.3E23)	(3.4E-6, 8.3E-5, 1.1E23)
HepPh	(4.7E-5, 8.8E-4, 8.5E17)	(5.1E-5, 9.5E-4, 6.7E17)	(5.5E-5, 1.0E-3, 5.3E17)	(5.7E-5, 1.1E-3, 4.6E17)	(6.2E-5, 1.2E-3, 3.7E17)
AstroPh	(3.0E-5, 5.9E-4, 5.2E18)	(3.1E-5, 6.1E-4, 4.6E18)	(3.4E-5, 6.6E-4, 3.7E18)	(3.5E-5, 6.9E-4, 3.1E18)	(3.7E-5, 7.3E-4, 2.6E18)
CondMat	(2.6E-5, 5.2E-4, 2.5E19)	(2.6E-5, 5.2E-4, 2.5E19)	(2.8E-5, 5.6E-4, 2.0E19)	(3.0E-5, 6.0E-4, 1.7E19)	(3.2E-5, 6.3E-4, 1.4E19)
DBLP	(1.7E-6, 4.3E-5, 2.2E24)	(1.9E-6, 4.8E-5, 1.6E24)	(1.9E-6, 4.8E-5, 1.6E24)	(2.1E-6, 5.3E-5, 1.2E24)	(2.2E-6, 5.7E-5, 9.5E23)
Enron	(1.7E-5, 3.6E-4, 1.1E20)	(1.7E-5, 3.6E-4, 1.1E20)	(1.8E-5, 3.8E-4, 9.3E19)	(2.0E-5, 4.2E-4, 7.1E19)	(2.0E-5, 4.2E-4, 7.1E19)
EuAll	(3.8E-6, 9.4E-5, 2.9E23)	(3.8E-6, 9.4E-5, 2.9E23)	(3.8E-6, 9.4E-5, 2.9E23)	(3.8E-6, 9.4E-5, 2.9E23)	(3.8E-6, 9.4E-5, 2.9E23)
Wiki	(3.3E-7, 9.7E-6, 4.3E27)	(3.3E-7, 9.7E-6, 4.3E27)	(3.3E-7, 9.7E-6, 4.3E27)	(3.3E-7, 9.7E-6, 4.3E27)	(3.3E-7, 9.7E-6, 4.3E27)
AS733	(9.4E-5, 1.7E-3, 2.9E17)	(9.4E-5, 1.7E-3, 2.9E17)	(9.4E-5, 1.7E-3, 2.9E17)	(1.1E-4, 2.0E-3, 1.6E17)	(1.1E-4, 2.0E-3, 1.6E17)
Oregon	(5.1E-5, 9.5E-4, 2.6E18)	(5.1E-5, 9.5E-4, 2.6E18)	(6.7E-5, 1.3E-3, 1.1E18)	(6.7E-5, 1.3E-3, 1.1E18)	(6.7E-5, 1.3E-3, 1.1E18)
Caida	(2.3E-5, 4.7E-4, 9.6E19)	(2.3E-5, 4.7E-4, 9.6E19)	(2.3E-5, 4.7E-4, 9.6E19)	(3.1E-5, 6.3E-4, 4.1E19)	(3.1E-5, 6.3E-4, 4.1E19)
Skitter	(3.2E-7, 9.0E-6, 1.0E27)	(3.4E-7, 9.8E-6, 8.0E26)	(3.7E-7, 1.1E-5, 6.5E26)	(3.9E-7, 1.1E-5, 5.4E26)	(4.1E-7, 1.2E-5, 4.7E26)
Gnutella3	(2.4E-5, 4.8E-4, 7.3E19)	(2.4E-5, 4.8E-4, 7.3E19)	(2.4E-5, 4.8E-4, 7.3E19)	(2.4E-5, 4.8E-4, 7.3E19)	(2.6E-5, 5.3E-4, 5.4E19)
Gnutella4	(1.8E-5, 3.7E-4, 2.6E20)	(1.8E-5, 3.7E-4, 2.6E20)	(1.8E-5, 3.7E-4, 2.6E20)	(1.8E-5, 3.7E-4, 2.6E20)	(1.9E-5, 4.1E-4, 2.0E20)
Gnutella5	(1.0E-5, 2.3E-4, 2.3E21)	(1.0E-5, 2.3E-4, 2.3E21)	(1.0E-5, 2.3E-4, 2.3E21)	(1.0E-5, 2.3E-4, 2.3E21)	(1.2E-5, 2.5E-4, 1.7E21)

Finally, we also want to evaluate the required bounds on $(\Omega(\wp), \Omega(f_{\mathbf{D}}), \Omega(n))$ in $(1-\epsilon)$-perfect DA. We demonstrate the results in Tab. 5 and make the following observations.

(i) Theoretically, the condition on the lower bound of $\wp$ is very loose, e.g., when $\epsilon = .1$, $\Omega(\wp) = 1.1\text{E-}7$ for Google+ and $\Omega(\wp) = 1.7\text{E-}7$ for Orkut, which suggests that $(1-\epsilon)$-perfect DA is implementable in practice. On the other hand, we can also see that the theoretical loose requirement on $\Omega(\wp)$ is at the expense of a strong condition on $\Omega(n)$, e.g., when $\epsilon = .1$, $\Omega(n) = 2.2\text{E}28$ for Google+ and $\Omega(n) = 2.0\text{E}27$ for Orkut. Consequently, to de-anonymize most of existing structural datasets which have sizes of million-level or less, a higher $\wp$ is desired (as we show in Tab. 2, 3, and 4).

(ii) From Tab. 5, we can see that the conditions on $\Omega(f_{\mathbf{D}})$ and $\Omega(n)$ exhibit the same behavior as in perfect DA, i.e., $\Omega(f_{\mathbf{D}})$ increases and $\Omega(n)$ decreases as $\Omega(\wp)$ increases, which is consistent with our quantification. Again, this is because $f_{\mathbf{D}}$ is an increasing function of $\wp$ given $p_{\mathbf{D}}$ and $\Omega(n)$ decreases when more similarity appears between G^a and G^u.

(iii) From Tab. 5, we can also see that the impact of graph density on $\Omega(f_{\mathbf{D}})$ and $\Omega(n)$ is also similar to that in the perfect DA scenario.

6. OPTIMIZATION BASED DA PRACTICE

In Section 4, we comprehensively quantified conditions for perfect DA and $(1-\epsilon)$-perfect DA. Based on our large-scale study on 26 real world datasets in Section 5, we find most, if not all, existing structural datasets are de-anonymizable partially or completely (Tab. 3). Interestingly, our DA quantification leads to a DA scheme, denoted by $\mathfrak{A}^*$, straightforwardly. Basically, $\mathfrak{A}^*$ can be implemented as follows: we can calculate the DE caused by each σ_k $(1 \le k \le n!)$ and let σ_0 be the σ_k that induces the least DE. According to the quantification, the σ_0 produced by $\mathfrak{A}^*$ should be the optimum DA scheme. However, $\mathfrak{A}^*$ is computationally infeasible in practice due to its high computational complexity $O(n!)$. In this section, we present a novel relaxed and operational version of $\mathfrak{A}^*$ followed by analyzing its performance theoretically and experimentally on large scale real datasets.

6.1 Optimization based DA

Before proposing our relaxed and computationally feasible version of $\mathfrak{A}^*$, we define some useful *structural features* for $i \in V^a$ or V^u as follows.

Degree: For $i \in V^a$ (resp., V^u), its *degree feature* $f_d(i)$ is its degree in G^a (resp., G^u), i.e., $f_d(i) = |N_i^a|$ (resp., $|N_i^u|$).

Neighborhood: For $i \in V^a$ (resp., V^u), its *neighborhood feature* $\overline{f_n(i)}$ is a *β-dimensional vector* $(d_1^i, d_2^i, \cdots, d_\beta^i)$, where d_k^i $(1 \le k \le \beta)$ is the k-th largest degree in $\{|N_j^a||j \in N_i^a\}$ (resp., $\{|N_j^u||j \in N_i^u\}$), i.e., d_k^i is the k-th largest degree of the neighboring users of i. In the case that $|N_i^a| < \beta$ (resp., $|N_i^u| < \beta$), we set $d^i_{|N_i^a|+1} = d^i_{|N_i^a|+2} = \cdots = d^i_\beta = \Delta^a$ (resp., $d^i_{|N_i^u|+1} = d^i_{|N_i^u|+2} = \cdots = d^i_\beta = \Delta^u$), where $\Delta^a = \max\{|N_i^a||i \in V^a\}$ (resp., $\Delta^u = \max\{|N_i^u||i \in V^u\}$) is the maximum degree of G^a (resp., G^u).

Top-K reference distance: For $i \in V^a$ (resp., V^u), its *Top-K reference distance feature* $\overline{f_K(i)}$ is a *K-dimensional vector* $(h_1^i, h_2^i, \cdots, h_K^i)$, where h_k^i $(1 \le k \le K)$ is the distance (the length of a shortest path) from i to the user with the k-th largest degree in G^a (resp., G^u). Note that it is possible $h_k^i = \infty$ if the graph is not connected.

Landmark reference distance: Suppose $V_L^a = \{v_1, v_2, \cdots, v_L | v_k \in V^a\}$ is a set of users that has been de-anonymized (evidently, $V_L^a = \emptyset$ initially) to $U_L^u = \{u_1, u_2, \cdots, u_L | u_k \in V^u\}$ under some DA scheme σ with $\sigma(v_k) = u_k$ $(1 \le k \le L)$. Intuitively, V_L^a and U_L^u can be used as auxiliary information for future DA. Therefore, for $i \in V^a \setminus V_L^a$ (resp., $V^u \setminus U_L^u$), we define its *landmark reference distance feature* $\overline{f_l(i)} = (h_1^i, h_2^i, \cdots, h_L^i)$, where h_k^i $(1 \le k \le L)$ is the distance from i to $v_k \in V_L^a$ (resp., $u_k \in U_L^u$).

Sampling closeness centrality: For $i \in V^a$ (resp., V^u), we define the *sampling closeness centrality feature* $f_c(i)$ to characterize its global topological property without inducing too much computational overhead. Formally, we first randomly

sample a subset S^a of V^a (resp., S^u of V^u). Then, we define $f_c(i) = \sum_{j \in S^a \setminus \{i\}} \frac{1}{h(i,j)}$ (resp., $f_c(i) = \sum_{j \in S^u \setminus \{i\}} \frac{1}{h(i,j)}$), where $h(i,j)$ is the distance from i to j.

According to the aforementioned definitions, (i) we consider both local and global structural features of a user, e.g., the degree and neighborhood features characterize the local topological properties of a user while the Top-K reference distance and sampling closeness centrality features demonstrate the global topological characteristics of a user; (ii) we also consider the computational efficiency of obtaining these features for a user. For instance, instead of using the accurate *closeness centrality*, we introduce a sampling closeness centrality feature, which can characterize the global feature of a user without causing too much computation overhead.

Now, based on the features defined for each user, we can quantitatively measure the *similarity* between an anonymized user $i \in V^a$ and a known user $j \in V^u$. Let $\overline{f_{d,c}(i)} = (f_d(i), f_c(i))$. Then, we define the *structural similarity* between $i \in V^a$ and $j \in V^u$ as $\phi(i,j) = c_1 \cdot s(\overline{f_{d,c}(i)}, \overline{f_{d,c}(j)}) + c_2 \cdot s(\overline{f_n(i)}, \overline{f_n(j)}) + c_3 \cdot s(\overline{f_K(i)}, \overline{f_K(j)}) + c_4 \cdot s(\overline{f_l(i)}, \overline{f_l(j)})$, where $c_{1,2,3,4} \in [0,1]$ are constant values representing the weights and $c_1 + c_2 + c_3 + c_4 = 1$, and $s(\cdot,\cdot)$ is the *Cosine similarity* between two vectors.

According to our theoretical quantification in Section 4, $\mathfrak{A}^*$ is inherently an optimization based algorithm with the objective of minimizing the DE Ψ_{σ_k}, which is different from most of existing DA algorithms (heuristics based) [1][2][3]. Inspired by our quantification, we design a novel and operational ***O**ptimization based **D**e-**A**nonymization* (**ODA**) scheme, which is a relaxed version of $\mathfrak{A}^*$. In ODA, rather than using the DE function as in the quantification, we re-define $\psi_{i,j}$ and Ψ_σ as follows. Given a DA scheme $\sigma = \{(i,j) | i \in V^a, j \in V^u\}$, we define the DE on a user mapping $(i,j) \in \sigma$ as $\psi_{i,j} = |f_d(i) - f_d(j)| + (1 - \phi(i,j)) \cdot |f_d(i) - f_d(j)|$, and the DE on σ as $\Psi_\sigma = \sum_{(i,j) \in \sigma} \psi_{i,j}$. Based on Ψ_σ, we give the framework of ODA as shown in Algorithm 1. In Algorithm 1, $\Lambda^a \subseteq V^a$ is the target DA set and $\Lambda^u \subseteq V^u$ is the possible mapping set of Λ^a. GetTopDegree(X, y) is a function to return y users with the largest degree values in X, i.e., return $\{i | i$ has the Top-y degree in $X\}$. $\mathcal{C}(i) \subseteq \Lambda^u$ is the *candidate mapping set* for i, which consists of the γ most possible mappings of i in Λ^u. GetTopSimilarity(i, Λ^u, γ) is a function to return γ users having the highest similarity scores with i in Λ^u, i.e., return $\{j | j \in \Lambda^u$, and j has the Top-γ $\phi(i,j)$ in $\Lambda^u\}$.

From Algorithm 1, ODA de-anonymizes G^a iteratively. During each iteration, ODA is trying to de-anonymize a subset of V^a and seeking the *sub-DA scheme* $\sigma^*(\Lambda^a)$ which induces the least DE. We explain the idea of ODA in details as follows. In Line 3, we initialize the target DA set Λ^a and the candidate mapping set Λ^u. From the initialization, α is an important parameter to control how many anonymized users will be processed in each iteration. In Line 4, we compute a *candidate mapping set* $\mathcal{C}(i)$ for each $i \in \Lambda^a$. $\mathcal{C}(i)$ consists γ most similar users of i in Λ^u. Here, we define $\mathcal{C}(\cdot)$ mainly for reducing the computational complexity. Instead of trying every mapping from i to Λ^u, we only consider to map i to some user in $\mathcal{C}(i)$. Hence, γ is another important parameter to control the computational complexity of ODA. We will demonstrate how to set α and γ to make ODA computationally feasible in Theorem 9. In Line 5, we find a DA scheme $\sigma^*(\Lambda^a)$ on Λ^a such that $\Psi_{\sigma^*(\Lambda^a)} = \min\{\Psi_{\sigma(\Lambda^a)} | \sigma(\Lambda^a) \in \prod_{i \in \Lambda^a} (i \times \mathcal{C}(i))\}$, i.e., $\sigma^*(\Lambda^a)$ causes the least DE. Furthermore, the *consistent rule* and the *pruning rule* are applied in this step. The *consistent rule* makes any possible DA scheme $\sigma(\Lambda^a)$ consistent, i.e., no *mapping confliction* which is defined as the situation that two or more anonymized users are mapped to the same known user. This is because it is possible that $\mathcal{C}(i_1) \cap \mathcal{C}(i_2) \neq \emptyset$ for $i_1 \neq i_2 \in \Lambda^a$, and the situation $\sigma(i_1) = \sigma(i_2)$ in a DA scheme should be avoided. Note that, it is possible that no $\sigma(\Lambda^a)$ is consistent. In this case, we should increase γ to guarantee at least one $\sigma(\Lambda^a)$ is consistent. The *pruning rule* is used to remove some DA schemes whose DE is larger than the current known least DE. For instance, let $\sigma^*(\Lambda^a)$ be the DA scheme having the least DE after testing k possible DA schemes. Then, when testing the $(k+1)$-th possible DA scheme $\sigma_{k+1}(\Lambda^a)$, if partial of mappings in $\sigma_{k+1}(\Lambda^a)$ has already induced a larger DE than $\sigma^*(\Lambda^a)$, we stop test $\sigma_{k+1}(\Lambda^a)$ and continue the next one. On the other hand, if $\sigma_{k+1}(\Lambda^a)$ induces a smaller DE than $\sigma^*(\Lambda^a)$, we update $\sigma^*(\Lambda^a)$ to $\sigma_{k+1}(\Lambda^a)$. Both the consistent rule and the pruning rule can remove some unqualified DA schemes in advance, which can speed up ODA. Actually, although $\sigma^*(\Lambda^a)$ causes the least DE, $\sigma^*(\Lambda^a)$ is a local optimization solution (according to our quantification, $\mathfrak{A}^*$ induces the optimum solution). This is because we try to seek a tradeoff between computational feasibility and DA accuracy. After obtaining $\sigma^*(\Lambda^a)$, we accept the mappings in $\sigma^*(\Lambda^a)$ with similarity scores no less than a *threshold value* θ (Lines 6-8). For the mappings been rejected, they will be re-considered in the following iterations for possible better DAs. If no mapping can be accepted, we stop ODA. Finally, we analyze the time and space complexities of ODA in Theorem 9. We defer the proof to Appendix C for readability.

Algorithm 1: ODA

```
1 Define Λ^a = Λ^u = ∅;
2 while true do
3     Λ^a = GetTopDegree(V^a, α), Λ^u = GetTopDegree(V^u, α);
4     for every i ∈ Λ^a, compute a candidate mapping set C(i) = GetTopSimilarity(i, Λ^u, γ);
5     apply the consistent rule and pruning rule to find the DA scheme σ(Λ^a) ∈ ∏_{i∈Λ^a} (i × C(i)) which induces the least DE Ψ_{σ(Λ^a)}, denoted by σ*(Λ^a) = {(i_1, j_1), (i_2, j_2), ··· , (i_α, j_α)};
6     for each (i, j) ∈ σ*(Λ^a), if φ(i, j) ≥ θ then
7         accept the mapping (i, j);
8         V^a = V^a \ {i}, V^u = V^u \ {j};
9     if no mapping in σ*(Λ^a) is accepted, break;
```

THEOREM 9. *(i) The space complexity of ODA is* $O(\min\{n^2, m+n\})$. *(ii) Let* γ *be some constant value,* $\alpha = \Theta(\log n)$, *and* Γ *be the average number of accepted mappings in each iteration of ODA. Then, the time complexity of ODA is* $O(m + n \log n + n^{\Theta(1)\log\gamma + 1}/\Gamma)$ *in the worst case.*

Finally, we make some remarks on ODA as follows.

(i) ODA is a *cold start* algorithm, i.e., we do not need any priori knowledge, e.g., the seed mapping information

[1][2][3], to bootstrap the DA process. Furthermore, unlike existing DA algorithms [1][2][3] which consist of two phases (*landmark identification phase* and *DA propagation phase*), ODA is a single-phase algorithm. Interestingly, ODA itself can act as a *landmark identification algorithm*. From our experiment (Section 6.2), ODA can de-anonymize the 60-180 Top-degree users in Gowalla and Google+ (see Tab. 1) perfectly, which can serve as landmarks (V_L^a and U_L^u) for future DAs. In addition, ODA as a landmark identification algorithm is much faster than that in [2] (with complexity of $O(nd^{k-1}) = O(n^k)$, where d is maximum degree of G^a/G^u and k is the number of landmarks) and [3] (with complexity of $k!$, could be computationally infeasible for a PC when $k \geq 20$).

(*ii*) Similar to $\mathfrak{A}^*$, ODA is an optimization based DA scheme, which is different from most of existing heuristics based solutions [1][2][3]. In ODA, the objective is to minimize a DE function. The reasonableness and soundness of ODA lie on one direct conclusion of our theoretical quantification: *minimizing the DE leads to the best possible DA scheme*.

(*iii*) In ODA, we seek an adjustable tradeoff between DA accuracy and computational feasibility. Although $\mathfrak{A}^*$ obtains the optimum solution *a.a.s.* in terms of our quantification, it is computationally infeasible ($O(n!)$). ODA has a polynomial time complexity of $O(m + n\log n + n^{\Theta(1)\log\gamma+1}/\Gamma)$ in the worst case, which is computationally feasible at the cost of sacrificing some accuracy. Based on our experiments on large scale real datasets in the following subsection, ODA is operable while preserves satisfiable DA performance.

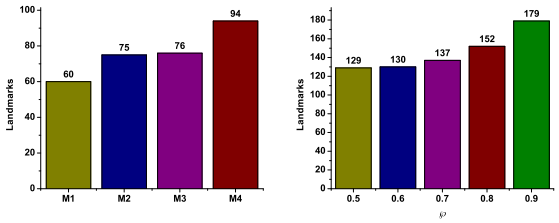
(*iv*) ODA is a general framework. Line 5 can also be implemented by seeking a *maximum weighted bipartite graph matching* on a *weighted bipartite graph* $G(\Lambda^a \cup \Lambda^u, \bigcup_{i\in\Lambda^a}(i \times \mathcal{C}(i)))$, where the weight on each edge is $\phi(i,j)$ $(i \in \Lambda^a, j \in \mathcal{C}(i))$.

(*v*) In ODA, one implicit assumption is $V^a = V^u$, i.e., the G^a and G^u are defined on the same group of users. In practice, it is possible that V^a and V^u are not exactly the same. In this case, if V^a and V^u are not significantly different, ODA is also workable at the cost of some performance degradation ($(1-\epsilon)$-perfect DA). One better solution could be estimating the overlap between G^a and G^u first, and then apply ODA to the overlap. We will take the estimation of the overlap between G^a and G^u as one of the future works.

6.2 Experimental Evaluation and Analysis

6.2.1 Datasets and Setup

We evaluate the performance of ODA on two real world datasets: Gowalla and Google+ (see the basic information in Section 5). Gowalla is a location based SN and consists of two different datasets [26][27]. The first dataset is a spatiotemporal mobility trace consisting of 6.44M *check-ins* generated by .2M users. Each check-in has the format of <UserID, latitude, longitude, timestamp, location ID>. The second dataset is a social graph (1M edges) of the same .2M users. Assume the mobility trace is anonymized. Our objective is to de-anonymize the mobility trace using the social graph as auxiliary data. Since the mobility trace does not have an explicit graph structure, supposing the social graph is the ground truth, we apply the technique in [27] on the mobility trace to construct four graphs with different *recalls* and *precisions*, denoted by $M1, M2, M3$, and $M4$, respectively (recall $= \frac{\text{tp}}{\text{tp+fn}}$ and precision $= \frac{\text{tp}}{\text{tp+fp}}$, where tp = true positive, fp = false positive, and fn = false negative). Particularly, the recall and precision of $M1$ are 0.6 and 0.865, of $M2$ are 0.72 and 0.83, of $M3$ are 0.75 and 0.78, and of $M4$ are 0.8 and 0.72, respectively. The second dataset considered is the Google+ dataset in Section 5, which has 4.7M users and 90.8M edges. Given some projection probability $\wp \in [0.5, 0.9]$, We first use the *projection process* in Section 4 to produce G^a and G^u, and then use ODA to de-anonymize G^a with G^u as auxiliary data. Note that, the auxiliary data is from a different domain (social data) with the anonymized data (mobility trace) in Gowalla while the auxiliary and anonymized data are from the same domain in Google+.



(a) Gowalla (b) Google+

Figure 2: Landmark identification. $c_1, c_2 \in [0.1, 0.3], c_3 \in [0.4, 0.8], c_4 = 0, \alpha \in [10, 30], \gamma \in [1, 4]$.

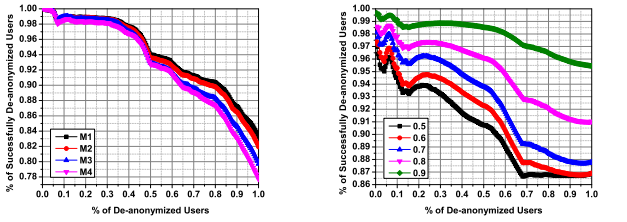
All the experiments are implemented on a PC with 64 bit Ubuntu 12.04 LTS OS, Intel Xeon E5620 CPU (2.4GHz × 8 Threads), 48GB memory, and 2 disks with 8TB storage.

6.2.2 Results

Landmark Identification. As we mentioned in the previous subsection, ODA itself can work as a *landmark identification algorithm*. Let $V_L^a = U_L^u = \emptyset$ in ODA, i.e., $s(\overline{f_l(\cdot)}, \overline{f_l(\cdot)}) = 0$ in $\phi(\cdot,\cdot)$. Then, we run ODA on Gowalla and Google+ to identify some landmarks as shown in Fig. 2 (note that, the DA in ODA is conducted according to the degree non-increasing order). The results show that we can de-anonymize the first 60-94 users in Gowalla and the first 129-179 users in Google+ perfectly (100% correctly). For instance, when $G^a = M2$ in Gowalla, the first 75 users are perfectly de-anonymizable and when $\wp = 0.7$, the first 137 users in Google+ are perfectly de-anonymizable. According to ODA, the identified landmarks can serve as references for future DA.

From Fig. 2 (a), when the recall increases, there are more common edges between G^a and G^u, which implies it is easier to identify the high degree users based on the increased structural information and thus more landmarks can be identified. Because of a similar reason, we can see from Fig. 2 (b) that more landmarks can be identified in Google+ for large $\wp$ due to more edge overlap between G^a and G^u.

DA Results. By taking the users identified in Fig. 2 as landmarks, we employ ODA to de-anonymize Gowalla ($M1$, $M2$, $M3$, $M4$) and Google+ (G^a with different $\wp$) as shown in Fig. 3, where the x-axis represents the *accumulated percentage of users de-anonymized* and the y-axis represents the *accumulated percentage of users successfully de-anonymized*. From Fig. 3, we can see that the successful DA rate is higher for large-degree users than that of



(a) De-anonymize Gowalla (b) De-anonymize Google+

Figure 3: De-anonymize Gowalla and Google+. $c_1, c_2 \in [0, 0.2], c_3 + c_4 \in [0.4, 1], \alpha \in [10, 30], \gamma \in [2, 10]$.

small-degree users, i.e., when x increases, the percentage of successfully deanonymized users generally show a decreasing trend. The reason is that large-degree users carry more structural information, which can thus be more accurately de-anonymizable. This can also be seen from our quantification. For Gowalla, we observe from Fig. 3(a) that although recall dominates the landmark identification process, the large-scale DA performance is impacted more by precision. Generally, a high precision implies this dataset is more de-anonymizable, e.g. $M4$. This is because a high precision implies a low false positive, which can be viewed as *noise* in practice, and thus the DA accuracy is better. For Google+, we see from Fig. 3 (b) that the G^a projected with a large $\wp$, e.g., $\wp = 0.9$, is more de-anonymizable. As shown in our quantification, this is because a large $\wp$ implies more similarity between G^a and G^u and thus more users can be successfully de-anonymized.

From Fig. 3, we also see that the DA performance of O-DA on Gowalla and Google+ is better than the evaluation results shown in Tab. 3, e.g., when $\wp = 0.9$, Tab. 3 indicates 91.2% of the users in Google+ are *a.a.s.* de-anonymizable while ODA successfully de-anonymizes 95.5% of the users. This is because the values shown in Tab. 3 are the lower bounds on de-anonymizable users. In summary, about $77.7\% - 83.3\%$ of the users in Gowalla and $86.9\% - 95.5\%$ of the users in Google+ are de-anonymizable. Thus, structure based DA is implementable and powerful in practice.

Time Consumption. We calculate the time consumption on de-anonymizing Gowalla and Google+. On average, the initialization time (used for initializations), execution time (used for executing the iterations in ODA), and total time are 1.79 mins, 1.6 mins, and 3.39 mins for Gowalla and 0.88 hours, 5.61 hours, and 6.49 hours for Google+, respectively.

7. IMPLICATIONS AND DISCUSSION

Based on our DA quantification, evaluation on real world datasets, and our implemented DA scheme ODA, we provide some implications of this paper in this section. We also discuss the impacts of our findings to *secure data publishing* in practice and provide guidelines for future data publishing.

Structural information may induce privacy leakage. Although we have some previous work that show structure based DA is possible, in this paper, we theoretically demonstrate the reasons by providing rigorous quantification under a general data model. From the quantification, structural information can enable large-scale perfect or $(1-\epsilon)$-perfect DA. Therefore, *for secure data publication, besides the data itself, the information carried by the data's structure is also essential and deserves dedicated consideration before released.*

The fact is that we still have a long way to go to achieve secure data publishing. From our large scale study on 26 real world datasets, most of the existing structural datasets are de-anonymizable based only on their structural information. On the other hand, existing anonymization techniques are vulnerable to structure based DA attacks. Therefore, *new anonymization techniques should be developed.* Meanwhile, since structural data release, sharing, or transferring has significant business and social value, *the data utility should be preserved in the new developed anonymization schemes.* In summary, *new secure data publishing schemes that properly achieve a balance between data privacy protection and data utility preservation must be developed.*

Suggestions for secure data publishing. Secure data publishing is important for businesses, research, and the society. However, with the wide availability of rich auxiliary information, especially with the emergence of *Collaborative Information Seeking* (CIS) systems and *data/knowledge brokers* [28][29], the privacy of people, businesses, governments, etc. will increasingly be compromised. For secure data publishing, some general suggestions are as follows. (i) *Carefully share data with or transfer data to third parties and partners.* Before sharing the data, the data owners should examine the dedicated applications to see if the data sharing is necessary. Based on the requirements of applications, the data could be shared in different granularity levels: *digest level*: share/transfer a digest/summary of the data to third parties or partners; *partial and density-controlled level*: based on our quantification, controlling the graph density could increase the difficulty of DA. Therefore, in this level, only a density-controlled anonymized version (e.g., by sampling) of a subset of the data (e.g., a community) is shared/transferred; *density-controlled level*: a density-controlled anonymized version of the data is shared or transferred; *full level*: an anonymized version of the full dataset is shared or transferred. (ii) *Evaluate the potential vulnerability of the dataset before actual publishing.* Before actually publishing the data, the data owners can evaluate the vulnerability of the data. For instance, the data (structural) can be evaluated using our quantification as in Section 5. (iii) *Develop proper policy on data collection.* Many structural data owners allow public data collection, e.g., Twitter, Facebook allow crawlers and other automatic programs to collect users' information online. This could increase the data DA risk by providing auxiliary information to adversaries. Therefore, it is better for data owners to develop proper policies to limit such public data collection.

8. CONCLUSION AND FUTURE WORK

In this paper, we study the quantification, practice, and implications of structural data DA. First, for the first time, we address several fundamental open problems in data DA research by quantifying the conditions for *perfect DA and* $(1-\epsilon)$*-perfect DA under a general data model.* This bridges the gap between structural data DA practice and theory. Second, we conduct a large scale study on the de-anonymizability of 26 diverse real world structural datasets, which turn out to be de-anonymizable partially or perfectly. Third, follow-

ing our quantification, we propose a *cold start single-phase Optimization based DA* (ODA) attack. We also analyze O-DA theoretically and experimentally. The experimental results show that $77.7\% - 83.3\%$ of the users in Gowalla (.2M users, 1M edges) and $86.9\% - 95.5\%$ of the users in Google+ (4.7M users, 90.8M edges) can be de-anonymized, which implies structure based DA is implementable and powerful in practice. Finally, we conclude with some implications from our findings and provide some general suggestions for future secure data publishing.

Our future work will focus on the following: (i) We will evaluate our quantification on more structural datasets to further examine its generality. We also plan to improve O-DA to make it more efficient and robust; (ii) Since existing anonymization techniques are vulnerable to structure based DA attacks, we propose to develop application based effective schemes against such attacks; (iii) Data utility is another important concern. We plan to study how to quantify the tradeoff between privacy and utility followed by proposing privacy protection schemes with utility preservation; and (iv) Finally, due to the importance of secure data publishing, we propose to develop a *secure data publishing platform* in the future, which is expected to be invulnerable to both semantics based and structure based DA attacks.

Acknowledgments

The authors are very grateful to Nana Li and Jing S. He for helpful discussions on graph theory, to Huy Pham who helped us to process the Gowalla mobility trace (Huy Pham also shared a social strength graph obtained from the mobility trace of Gowalla users at Texas), and to Neil Z. Gong who shared the Google+ dataset with us.

This work was partly supported by NSF-CAREER-CNS-0545667. Mudhakar Srivatsa's research was sponsored by US Army Research laboratory and the UK Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory, the U.S. Government, the UK Ministry of Defense, or the UK Government. The US and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

9. REFERENCES

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

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

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

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

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

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

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

[8] N. Li, W. Qardaji, and D. Su, *On Sampling, Anonymization, and Differential Privacy Or, K-Anonymization Meets Differential Privacy*, ASIACCS 2012.

[9] C. Dwork, *Differential Privacy*, ICALP 2006.

[10] A. Korolova, R. Motwani, S. U. Nabar, and Y. Xu, *Link Privacy in Social Networks*, CIKM 2008.

[11] E. Zheleva and L. Getoor, *To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles*, WWW 2009.

[12] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall, *802.11 User Fingerprinting*, Mobicom 2007.

[13] L. Backstrom, E. Sun, and C. Marlow, *Find me If You Can: Improving Geographical Prediction with Social and Spatial Proximity*, WWW 2010.

[14] S. Han, V. Liu, Q. Pu, S. Peter, T. Anderson, A. Krishnamurthy, and D. Wetherall, *Expressive Privacy Control with Pseudonyms*, Sigcomm 2013.

[15] P. Mittal, M. Wright, and N. Borisov, *Pisces: Anonymous Communication Using Social Networks*, NDSS 2013.

[16] J. Kannan, G. Altekar, P. Maniatis, and B.-G. Chun *Making programs forget: Enforcing Lifetime for Sensitive Data*, USENIX 2013.

[17] M. Egele, G. Stringhini, C. Krugel, and G. Vigna, *COMPA: Detecting Compromised Accounts on Social Networks*, NDSS 2013.

[18] K. Singh, S. Bhola, and W. Lee, *xBook: Redesigning Privacy Control in Social Networking Pl atforms*, USENIX 2009.

[19] R. Shokri, G. Theodorakopoulos, J.-Y. L. Boudec, and J.-P. Hubaux, *Quantifying Location Privacy*, S&P 2011.

[20] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. L. Boudec, *Protecting Location Privacy: Optimal Strategy against Localization Attacks*, CCS 2012.

[21] M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, 2010.

[22] M. E. J. Newman, *The Structure and Function of Complex Networks*, SIAM Review, No. 45, pp. 167-256, 2003.

[23] B. Bollobás, *Random Graphs (Second Edition)*, Cambridge University Press, 2001.

[24] J. Riordan, *An Introduction to Combinatorial Analysis*, Wiley, 1958.

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

[26] *http://snap.stanford.edu/data/*

[27] H. Pham, C. Shahabi, and Yan Liu, *EBM - An Entropy-Based Model to Infer Social Strength from Spatiotemporal Data*, SIGMOD 2013.

[28] C. Shah, R. Capra, and P. Hansen, *Collaborative Information Seeking*, Computer, 2014.

[29] Z. Xu, J. Ramanathan, and R. Ramnath, *Identifying Knowledge Brokers and Their Role in Enterprise Research through Social Media*, Computer, 2014.

APPENDIX

A. PROOF SKETCH OF THEOREM 1

Since k is the number of incorrect mappings in $\sigma \neq \sigma_0$, $2 \leq k \leq n$ is evidently. For convenience of proof, let σ_k be a DA scheme that has k incorrect mappings. Under σ_k, let $V_k \subseteq V$ be the set of incorrectly de-anonymized users[5], $\mathcal{E}_k = \{e_{i,j} | i \in V_k \text{ or } j \in V_k\}$ be the set of all possible edges adjacent to at least one user in V_k, $\mathcal{E}_\tau = \{e_{i,j} | i, j \in V_k, (i,j) \in \sigma_k, \text{ and } (j,i) \in \sigma_k\}$ be the set of all possible edges corresponding to *transposition mappings*[6] in σ_k, and $\mathcal{E} = \{e_{i,j} | 1 \leq i \neq j \leq n\}$ be the set of all possible edges on

[5]Without of causing any confusion, we use V, V^a, and V^u interchangeably since $V = V^a = V^u$.

[6]If both mappings (i,j) and (j,i) are in σ_k, then $\{(i,j),(j,i)\}$ is called a *transposition mapping*, i.e., two users are incorrectly de-anonymized to each other.

V. Furthermore, define $m_k = |\mathcal{E}_k|$ and $m_\tau = |\mathcal{E}_\tau|$. Then, we have $|V_k| = k$, $m_k = \binom{k}{2} + k(n-k)$, $m_\tau \leq \frac{k}{2}$ since there are at most $\frac{k}{2}$ transposition mappings in σ_k, $|\mathcal{E}| = \binom{n}{2}$, and $\forall e_{i,j} \in \mathcal{E}$, $\Pr(e_{i,j} \in E) = p_{i,j} = \frac{d_i d_j}{2m-1}$.

Now, we quantify Ψ_{σ_0} stochastically. Actually, to quantify Ψ_{σ_0}, we considering the DE caused by the projection of each edge rather than considering the mapping directly. $\forall e_{i,j} \in \mathcal{E}$, if it appears in E and is projected to either G^a or G^u but not both during the edge projection process, then according to the definition of DE, it will cause a DE of 2. Consequently, the DE caused by $e_{i,j}$ satisfies a binomial distribution $\mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp))$. Furthermore, since the projection process is *i.i.d.* and considering Lemma 1, we have $\Psi_{\sigma_0} = \sum_{(t,t') \in \sigma_0} \psi_{t,t'} \sim \sum_{e_{i,j} \in \mathcal{E}} \mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp)) = \mathbf{B}(\sum_{e_{i,j} \in \mathcal{E}} 2, 2p_{i,j} \cdot \wp(1-\wp))$.

When we quantify Ψ_{σ_k}, we consider three cases respectively. *Case 1*: for $\forall e_{i,j} \in \mathcal{E} \setminus \mathcal{E}_k$, the DE caused by $e_{i,j}$ during the projection process also satisfies the binomial distribution $\mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp))$ since $i, j \in V \setminus V_k$ (i.e., i, j are successfully de-anonymized under σ_k). *Case 2:* for $\forall e_{i,j} \in \mathcal{E}_k \setminus \mathcal{E}_\tau$, it will be mapped to some other possible edge $\sigma_k(e_{i,j}) = e_{\sigma_k(i),\sigma_k(j)} \in \mathcal{E}$ since $e_{i,j} \notin \mathcal{E}_\tau$ and at least one of i and j is incorrectly de-anonymized under σ_k. Therefore, in this case, the DE caused by $e_{i,j}$ during the projection process satisfies binomial distribution $\mathbf{B}(2, p_{i,j} \cdot \wp(1 - p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)} \cdot \wp(1 - p_{i,j}\wp))$. *Case 3:* for $\forall e_{i,j} \in \mathcal{E}_\tau$, since it corresponds to a transposition mapping, the DE caused by $e_{i,j}$ during the projection process also satisfies the binomial distribution $\mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp))$. In summary, we have $\Psi_{\sigma_k} = \sum_{(t,t') \in \sigma_k} \psi_{t,t'} \sim \sum_{e_{i,j} \in \mathcal{E} \setminus \mathcal{E}_k} \mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp)) + \sum_{e_{i,j} \in \mathcal{E}_k \setminus \mathcal{E}_\tau} \mathbf{B}(2, p_{i,j} \cdot \wp(1 - p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)} \cdot \wp(1 - p_{i,j}\wp)) + \sum_{e_{i,j} \in \mathcal{E}_\tau} \mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp)) \underset{stochastically}{\geq} \sum_{e_{i,j} \in \mathcal{E} \setminus \mathcal{E}_k} \mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp)) + \sum_{e_{i,j} \in \mathcal{E}_k \setminus \mathcal{E}_\tau} \mathbf{B}(2, p_{i,j} \cdot \wp(1 - p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)} \cdot \wp(1 - p_{i,j}\wp)) = \mathbf{B}(\sum_{e_{i,j} \in \mathcal{E} \setminus \mathcal{E}_k} 2, 2p_{i,j} \cdot \wp(1-\wp)) + \mathbf{B}(\sum_{e_{i,j} \in \mathcal{E}_k \setminus \mathcal{E}_\tau} 2, p_{i,j} \cdot \wp(1 - p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)} \cdot \wp(1 - p_{i,j}\wp))$.

Now, define $X \sim \mathbf{B}(\sum_{e_{i,j} \in \mathcal{E}_k \setminus \mathcal{E}_\tau} 2, p_{i,j} \cdot \wp(1 - p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)} \cdot \wp(1 - p_{i,j}\wp))$ and $Y \sim \mathbf{B}(\sum_{e_{i,j} \in \mathcal{E}_k} 2, 2p_{i,j} \cdot \wp(1-\wp))$. Let λ_x and λ_y by the mean values of X and Y, respectively. Thus, we have $\lambda_x = (\sum_{e_{i,j} \in \mathcal{E}_k \setminus \mathcal{E}_\tau} 2) \cdot [p_{i,j} \cdot \wp(1 - p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)} \cdot \wp(1 - p_{i,j}\wp)] \geq= 4l\wp(1 - h\wp)(m_k - m_\tau)$ and $\lambda_y \leq 4h\wp(1-\wp)m_k$. Then, $\forall \sigma_k$ ($k \in [2, n]$), $\Pr(\Psi_{\sigma_k} - \Psi_{\sigma_0} \leq 0) \underset{stochastically}{\leq} \Pr(X - Y \leq 0)$.

We now derive the upper bound on $\Pr(X - Y \leq 0)$. Since $\wp > \frac{h-l}{h-hl}$, $m_\tau \leq \frac{k}{2}$, and $m_k = \binom{k}{2} + k(n-k)$, $\wp > \frac{h-l}{h-hl} = \frac{(h-l)m_k}{(h-hl)m_k} \underset{n \to \infty}{\simeq} \frac{(h-l)m_k + lm_\tau}{(h-hl)m_k + lhm_\tau} \Rightarrow \lambda_x > \lambda_y$. Applying Lemma 1 and considering that $f_\wp = \Omega(\frac{2\ln n + 1}{kn})$, we have $\Pr(X - Y \leq 0) \leq 2\exp(-\frac{(\lambda_x - \lambda_y)^2}{8(\lambda_x + \lambda_y)}) \leq 2\exp(-f_\wp m_k)$ $= 2\exp(-\Omega(\frac{2\ln n+1}{kn}) \cdot (\binom{k}{2} + k(n-k))) \leq 2\exp(-2\ln n - 1) \leq \frac{1}{n^2}$.

Define $\zeta(2) = \sum_{n>0} \frac{1}{n^2}$. Then, $\zeta(2)$ is the *Euler-Riemann zeta function* with parameter 2 and thus $\zeta(2) = \frac{\pi^2}{6} < \infty$. Consequently, according to the *Borel-Cantelli Lemma*, it is *a.a.s.* that $X \geq Y$. It follows that it is *a.a.s.* that $\Psi_{\sigma_k} \geq \Psi_{\sigma_0}$ for $2 \leq k \leq n$, i.e., $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \to 1$ for any $\sigma \neq \sigma_0$. □

B. PROOF SKETCH OF THEOREM 2

Let $\mathbf{E}_\sigma$ be the event that $\Psi_\sigma \leq \Psi_{\sigma_0}$. Then, $\Pr(\mathbf{E}) = \Pr(\bigcup_\sigma \mathbf{E}_\sigma) = \Pr(\bigcup_{k=2}^{n} \bigcup_{\sigma_k} \mathbf{E}_{\sigma_k})$. Let ϱ_k be the number of de-anonymizatoin schemes having k incorrect mappings. Then, $\varrho_k = \binom{n}{k} \cdot !k \leq n^k$, where $!k$ is the subfactorial of k [24][5]. Then, considering that $f_\wp = \Omega(\frac{(k+3)\ln n + 1}{kn})$ and based on *Boole's inequality* and the proof of Theorem 1, we have $\Pr(\mathbf{E}) = \Pr(\bigcup_{k=2}^{n} \bigcup_{\sigma_k} \mathbf{E}_{\sigma_k}) \leq \sum_{k=2}^{n} \varrho_k \cdot \Pr(\Psi_{\sigma_k} \leq \Psi_{\sigma_0}) \leq \sum_{k=2}^{n} n^k \cdot 2\exp(-f_\wp m_k) \underset{n \to \infty}{\leq} \sum_{k=2}^{n} 2\exp(k\ln n - (k+3)\ln n - 1) \leq \sum_{k=2}^{n} \frac{1}{n^3} \leq \frac{1}{n^2}$. Again, since $\zeta(2) = \sum_{n>0} = \frac{\pi^2}{6} < \infty$, it is *a.a.s.* that $\Pr(\mathbf{E}) \to 0$ based on the *Borel-Cantelli Lemma*, i.e., it is *a.a.s.* that there exists no DA scheme such that $\sigma \neq \sigma_0$ and $\Psi_\sigma \leq \Psi_{\sigma_0}$. □

C. PROOF OF THEOREM 9

(*i*) ODA's space complexity is upper bound by $O(\min\{n^2, m+n\})$. The proof is straightforward and thus we omit it.

(*ii*) In ODA, we assume $f_d(i), \overline{f_n(i)}, \overline{f_K(i)}, \overline{f_l(i)}$, and $f_c(i)$ are computed before the iteration starts and the time consumption of computing these features is bounded by $O(m + n\log n)$. Then, from ODA, the worst case time consumption of each iteration is upper bounded by $\gamma^\alpha = \gamma^{\Theta(\log n)} = 2^{\log \gamma^{\Theta(\log n)}} = 2^{\Theta(\log n)\log\gamma} = n^{\Theta(1)\log\gamma}$. Furthermore, the number of iterations in ODA is $\Theta(n/\Gamma)$. It follows the worst case time complexity of ODA is $O(m + n\log n) + O(n^{\Theta(1)\log\gamma+1}/\Gamma) = O(m + n\log n + n^{\Theta(1)\log\gamma+1}/\Gamma)$. □