Structural Data De-Anonymization: Theory and Practice

Shouling Ji, *Member, IEEE, ACM*, Weiqing Li, *Student Member, IEEE, ACM*, Mudhakar Srivatsa, *Senior Member, IEEE*, and Raheem Beyah, *Senior Member, IEEE, ACM*

***Abstract*— In this paper, we study the quantification, practice, and implications of structural data de-anonymization, including social data, mobility traces, and so on. First, we answer several open questions in structural data de-anonymization by quantifying perfect and $(1-\epsilon)$-perfect structural data de-anonymization, where ϵ is the error tolerated by a de-anonymization scheme. To the best of our knowledge, this is the first work on quantifying structural data de-anonymization under a general data model, which closes the gap between the structural data de-anonymization practice and theory. Second, we conduct the first large-scale study on the de-anonymizability of 26 real world structural data sets, including social networks, collaborations networks, communication networks, autonomous systems, peer-to-peer networks, and so on. We also quantitatively show the perfect and $(1-\epsilon)$-perfect de-anonymization conditions of the 26 data sets. Third, following our quantification, we present a practical attack [a novel single-phase cold start optimization-based de-anonymization (ODA) algorithm]. An experimental analysis of ODA shows that ∼77.7%–83.3% of the users in Gowalla (196 591 users and 950 327 edges) and 86.9%–95.5% of the users in Google+ (4 692 671 users and 90 751 480 edges) are de-anonymizable in different scenarios, which implies that the structure-based de-anonymization is powerful in practice. Finally, we discuss the implications of our de-anonymization quantification and our ODA attack and provide some general suggestions for future secure data publishing.**

***Index Terms*— De-anonymization, quantification, graph data, structural data.**

I. Introduction

NOWADAYS, many data generated by computer networks and services have a graph structure, which is referred to as *graph/structural data*. For instance, it is straightforward to model social networks, network topologies, communication networks, etc. by graphs [3], [8], [9], [34]. Additionally, mobility traces (e.g., WiFi contacts, Instant Message contacts, Bluetooth contacts) can also be modeled as graphs (structural data) [4]. Even general spatiotemporal data (mobility traces) with the classical (*latitude*, *longitude*, *timestamp*) format can be transferred to structural data by applying sophisticated techniques [35]. On the other hand, since these structural data have huge commercial value to businesses and significant impacts to society [36], [37], the security and privacy issues that arise during data release to the public, sharing with commercial partners, or/and transferring to third parties are attracting increasing interest [2]–[4].

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 has been shown to be vulnerable to many de-anonymization attacks [9]–[11].[1] In parallel, some sophisticated anonymization schemes to protect structural data privacy, e.g., k-anonymity and its variants [9]–[11], were designed.[2] They can protect the privacy of structural data to some extent. However, they are susceptible to emerging *structure based de-anonymization attacks* due to the limitations of the schemes (e.g., they are syntactic properties based) and the rich amount of information available to adversaries [2]–[4] (see the detailed analysis in Section 1 of the *Supplementary File*).

In structure based de-anonymization 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 the attack powerful, e.g., using Flickr to de-anonymize Twitter [3], using Facebook to de-anonymize WiFi mobility traces [4]. Furthermore, the wide availability of auxiliary data makes the attack applicable and practical [3], [4].

Structure based de-anonymization attacks were initially presented in [2], where Backstrom et al. designed both active and passive attacks to break the privacy of social network

Manuscript received May 25, 2015; revised December 6, 2015; accepted January 28, 2016; approved by IEEE/ACM Transactions on Networking Editor D. Leith. The work of M. Srivatsa was supported in part by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Grant W911NF-06-3-0001. Preliminary results were presented at the ACM Conference on Computer and Communications Security (CCS) 2014 [1].

S. Ji is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China, and also with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: sji@gatech.edu).

W. Li, and R. Beyah are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: wli64@gatech.edu; rbeyah@ece.gatech.edu).

M. Srivatsa is with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: msrivats@us.ibm.com).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2016.2536479

[1] Intuitively, structural data can be modeled by graphs (see the data model in Section II). Within a graph, the structural correlation information (e.g., the combination of node degree, closeness centrality, betweenness centrality, relative distance to landmark users, and other graph topological properties) can be leveraged by adversaries to uniquely identify many users even if the PII is removed, and thus many users can be successfully de-anonymized [3]–[5].

[2] Note that, the *differential privacy* [12] is well developed to protect the privacy of *interactive data release*. However, it is difficult to apply differential privacy in its current form to defend against *structural data de-anonymization attacks* which are designed to breach the privacy of *non-interactive data release*. Detailed analysis can be found in [6].

users. However, since the attacks in [2] leverage the success of a "sybil" attack before actual anonymized data publication, they are difficult to extend to large scale datasets. Later, Narayanan and Shmatikov designed a new structure based de-anonymization attack in [3], which successfully de-anonymizes a large scale directed social network by applying several heuristics such as eccentricity, edge directionality, reverse match, etc. In [4], Srivatsa and Hicks demonstrated that the privacy of three kinds of mobility traces can be compromised by structure based de-anonymization attacks. However, the attacks presented in [4] are only suitable for small datasets due to its computational infeasibility on finding a proper landmark mappings for large datasets. Note that each of the aforementioned attacks consist of two phases: a *landmark identification phase* and a *de-anonymization propagation phase*.

Although we already have some successful structure based de-anonymization practices [2]–[4], we do not have any *rigorous theoretical results under a general model* yet in answering why structure based de-anonymization attacks work. In [8], 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., social networks, collaboration networks) do not follow the ER model. Actually, they may follow the *power-law model*, *exponential model*, etc. [29], [30], [34]. Therefore, under a practical *general data model*, there are still some open problems in de-anonymization research, including: (i) *why can structural data be de-anonymized*? (ii) *what are the conditions for successful structural data de-anonymization*? and (iii) *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 de-anonymization in this paper. Particularly, our contributions are as follows.

- To the best of our knowledge, this is the first work on quantifying structural data de-anonymization under a general data model. In our quantification, we answer several fundamental open problems: why structural data can be de-anonymized based only on the topological information (the inherent reason for the success of existing structure based de-anonymization practices)? what are the conditions for *perfect* and $(1-\epsilon)$*-perfect de-anonymization*, where ϵ is the *error* tolerated by a de-anonymization scheme? what portion of users can be de-anonymized in a structural dataset? Thus, we close the gap between structural data de-anonymization practice and theory.
- We conduct the first large-scale study on the de-anonymizability of 26 real world structural datasets, including social networks, location based mobility traces and social networks, collaboration networks, communication networks (Email, WikiTalk), autonomous system graph data, peer-to-peer network data, etc. Based on our study, we find *all* the considered structural datasets are de-anonymizable perfectly or partially. We also quantitatively show the conditions for perfect and $(1-\epsilon)$-perfect de-anonymization and what portion of users can be de-anonymized for the 26 datasets.
- Following our quantification, we present a novel *Optimization based De-Anonymization* (ODA) attack. Different from existing structure based de-anonymization attacks [2]–[4], 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 (196,591 users and 950,327 edges) and Google+ (4,692,671 users and 90,751,480 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 implies structure based de-anonymization is powerful in practice.
- Finally, we discuss some implications of this work according to our structural de-anonymization 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 II, we give the data and attack models. In Section III, we theoretically quantify perfect and $(1-\epsilon)$-perfect de-anonymization attacks under a general data model, followed by a large-scale evaluation of 26 diverse real world structural datasets in Section IV. In Section V, we present a novel optimization based de-anonymization attack with theoretical and experimental analysis. The paper is concluded and future work is addressed in Section VI. We summarize the related work and highlight the differences between this paper and existing works in Section I of the *Supplementary File*. We discuss the implications of our de-anonymization quantification and ODA attack in Section IV of the *Supplementary File*.

II. System Model

In this paper, we focus on quantifying and analyzing the de-anonymization attack (vulnerability) on anonymized structural data, which could be social data released by social network operators, e.g., Google+ [33], Facebook [34], Twitter [34], and/or mobility data generated by mobile devices, e.g., WiFi and Bluetooth traces [4], instant message contacts [4], email networks [34], classical longitude-latitude spatiotemporal traces [34], [35]. In the following subsection, we formally define the anonymized and auxiliary data models, as well as the attack model.

A. Data Model

It is straightforward to model social data using graphs, where nodes represent users and edges/links indicate the social relationships (*friendship*, *contact*, *following*) among users. For the mobility data generated by users (users' devices), they can also be modeled by contact graphs according to recently proposed techniques [4], [35]. Furthermore, it has been shown that a contact graph derived from mobility data has strong correlation (similarity) with the social graph of the same group of users that generated them [4], [35]. Therefore, we model the anonymized structural data by a graph $G^a = (V^a, E^a)$, where $V^a = \{i | i$ is an *anonymized* user$\}$ is the user set and $E^a = \{e^a_{i,j} |$ there is a relationship (friend, contact, etc.) between $i \in V^a$ 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 $N_i^a = \{j | \exists e_{i,j}^a \in E^a\}$ and we denote the cardinality of N_i^a as $|N_i^a|$, i.e., the degree of i.

The auxiliary data is also assumed to be structural data, e.g., a social network compromising users overlapped with that in the anonymized structural data [3], [4]. 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 [3], [4], [7], and [35]. Consequently, the auxiliary data is also modeled by a graph $G^u = (V^u, E^u)$, where $V^u = \{i$ is a *known* user$\}$ and $E^u = \{e_{i,j}^u |$ there is a relationship (friend, contact, etc.) between $i \in V^u$ and $j \in V^u\}$. Similarly, the neighborhood of $i \in V^u$ is defined as $N_i^u = \{j | \exists e_{i,j}^u \in E^u\}$.

B. De-Anonymization Attack

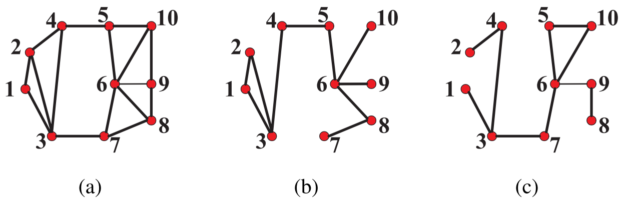
Given G^a and G^u, a de-anonymization attack can be formally defined as a *mapping*: $\sigma : V^a \to V^u$. For $\forall i \in V^a$, its mapping under σ is $\sigma(i) \in V^u$. Similarly, for $\forall e_{i,j}^a \in E^a$, $\sigma(e_{i,j}^a) = e_{\sigma(i),\sigma(j)}^u$. Note that, it is unknown whether the anonymized user i appears in the auxiliary dataset V^u or not in a practical de-anonymization attack. In the case that i does not appear in V^u, a correct de-anonymization of i is to map i to a special *not existing indicator* $\perp$. To avoid any confusion, mathematically, we assume that the *not existing indicator* $\perp$ is a *default* element of V^a and V^u. Under σ, a successful de-anonymization 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; if $\nexists i' \in V^u$ such that i and i' correspond to the same user, $\sigma(i) = \perp$. For other cases, the de-anonymization on i fails. Consequently, the objective of a de-anonymization attack is to successfully de-anonymize as many users in V^a as possible.

III. De-Anonymization Quantification

In this section, given G^a and G^u, we quantify a de-anonymization attack under an *arbitrary graph distribution* in multiple scenarios. Particularly, we study the condition on the structure of anonymized data under which a successful de-anonymization attack can be conducted. Note that, our quantification is aiming at providing a theoretical foundation on understanding the success of recent heuristic structure-based de-anonymization practices [3], [4]. We theoretically demonstrate that even without any further (e.g., semantic) knowledge, perfect or $(1-\epsilon)$-perfect de-anonymization attacks can be implemented when some structural conditions on the underlaying graph corresponding to G^a and G^u are satisfied.

A. 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[3] [3], [4], [8]. 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 overlap part, or (ii) redefine $V_{new}^a = V^a \cup (V^u \setminus V^a)$ and $V_{new}^u = V^u \cup (V^a \setminus V^u)$, 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 $G^a = (V_{new}^a, E^a)$ and $G^u = (V_{new}^u, E^u)$. To avoid confusion, we assume $V^a = V^u$ in the rest of this section.



(a) (b) (c)

Fig. 1. Edge/relationship projection. (a) G. (b) G^a. (c) G^u.

Second, similar to the methodology in [8], 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., "friendship" relationship on Facebook, "circle" relationship on Google+, "follow" relationship on Twitter, "co-occurrence" relationship in Gowalla, "coauthor" relationship in DBLP. The projection from G to G^a is characterized by an *edge/relationship projection process* [8]: (i) $V^a = V$; and (ii) $\forall e_{i,j} \in E$, it 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) 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 [3] and [4] also supports such strong structural correlation; (ii) it is straightforward to make this assumption more practical by further assuming that in addition to the projection process, some fake edges may be added to G^a and G^u with some probability. Our quantification can tackle this situation directly, however, with a more complicated expression when reporting the quantification results; and (iii) the assumption of an existing conceptual underlying graph G is only for the

[3] Note that, this assumption is only for our de-anonymization quantification. We do not make this assumption for a practical de-anonymization attack, e.g., the proposed ODA attack.

mathematical purpose of quantifying the structural correlation between G^a and G^u. Even without this assumption, the quantifications in this paper as well as that in [8] are still valid, however, they will be much more complicated since we need to introduce more functions to characterize the structural correlation between G^a and G^u.

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

B. 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 [8], Pedarsani and Grossglauser quantified the privacy of G when G is an ER random graph $G(n, p)$.[4] The $G(n, p)$ model is very useful as a source of insight into the study of structural data, e.g., social networks [8], [29]. 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., social networks, collaboration networks, mobility based contact networks) [29], [30]. Actually, the degree distribution of real world structural data (represented by graphs) may follow any distribution such as the power-law distribution and exponential distribution [29], [30]. Therefore, it is significant to understand and quantify a de-anonymization attack (or the privacy and vulnerability) for structural data under an *arbitrary degree distribution*. To this end, we characterize G by a generalized graph model, the *configuration model* [29]. Under the configuration model, a graph is specified by an arbitrary degree sequence $\mathbf{D}$ rather than a particular degree distribution.[5] Since $\mathbf{D}$ is an arbitrary degree sequence, $\mathbf{D}$ can follow an arbitrary degree distribution observed in real world data [29].

Let $p_{i,j}$ be the probability of existing an edge between $i, j \in V$. Then, we have $p_{i,j} = \frac{d_i d_j}{2m-1} \underset{\text{as } m\rightarrow\infty}{\simeq} \frac{d_i d_j}{2m}$, which is a key property of the configuration model [29]. 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 de-anonymization scheme $\sigma = \{(i, i') | 1 \leq i, i' \leq n, i \in V^a, i' \in V^u\}$, which is actually a subset of $V^a \times V^u$ (i.e., $\sigma \subseteq V^a \times V^u$), we define the *De-anonymization Error* (DE) on a user mapping $(i, i') \in \sigma$ as $\psi_{i,i'} = |N_i^a \setminus N_{i'}^u| + |N_{i'}^u \setminus N_i^a|$, which measures the neighborhoods' difference between i in G^a and i' in G^u under the particular σ.[6] Then, we define the overall DE for a particular σ as $\Psi_\sigma = \sum\limits_{(i,i')\in\sigma} \psi_{i,i'}$. Taking G^a and G^u shown in Fig. 1 as an example, the DE of the perfect de-anonymization scheme σ_0 is $\Psi_{\sigma_0} = 20$. For another de-anonymization 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; all the other users are correctly de-anonymized), its DE is $\Psi_\sigma = 28$.

In the following subsections, we quantify a de-anonymization attack by studying the conditions on G and the projection process under which perfect and $(1 - \epsilon)$-perfect de-anonymization attacks can be conducted. Equivalently, *we study the conditions on G and the projection process such that the perfect/$(1 - \epsilon)$-perfect de-anonymization scheme minimizes DE* (mathematically, this implies a perfect/$(1 - \epsilon)$-perfect de-anonymization scheme can be obtained *by minimizing the DE* since the number of de-anonymization schemes is bounded).

C. Perfect De-Anonymization

Now, we quantify the conditions for perfect de-anonymization attacks. To make the paper more readable, we place all the proofs in the *Supplementary File*.

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.*)[7] that $\Psi_\sigma \geq \Psi_{\sigma_0}$ for any de-anonymization scheme $\sigma \neq \sigma_0$.

Theorem 1: For any $\sigma \neq \sigma_0$, let k be the number of incorrect mappings in σ, i.e., $k = |\sigma \setminus \sigma_0|$. Then, $2 \leq k \leq n$ and $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \underset{n\rightarrow\infty}{\rightarrow} 1$ 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 de-anonymization scheme σ_0 will cause less DE than any other given de-anonymization scheme $\sigma \neq \sigma_0$. To guarantee the *uniqueness* of σ_0 (i.e., σ_0 is *the one and the only one* de-anonymization scheme introducing the least DE), intuitively, stronger conditions on $\wp, l$, and h are required. We quantify such conditions in Theorem 2.

Theorem 2: Let $\mathbf{E}$ be the event that there exists at least one de-anonymization 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 $2 \leq k \leq n$, $\Pr(\mathbf{E}) \rightarrow 0$, i.e., it is* a.a.s. *that there exists no de-anonymization scheme σ such that $\sigma \neq \sigma_0$ 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

[4] 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.

[5] Given a degree sequence $\mathbf{D} =< d_1, d_2, \cdots, d_n >$, a random graph with degree sequence $\mathbf{D}$ can be generated in the following manner [29]: give each node i a total of d_i "stubs". Then, there are $\sum_i d_i = 2m$ stubs in total, where m is the number of edges; randomly and uniformly choose two of the stubs and create an edge by connecting them; choose another pair from the remaining $2m - 2$ stubs, connect them, and so on until all the stubs are used up. More details and discussion can be found in [29].

[6] Note that, the DE can only be calculated after specifying a σ. Further, σ can be any de-anonymization scheme (i.e., not necessary to be the perfect de-anonymization scheme).

[7] We use the phrase *asymptotically almost surely* (*a.a.s.*) to denote an event that holds with probability tending to 1 as $n \rightarrow \infty$.

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 de-anonymization scheme causes the least DE. Furthermore, the number of possible de-anonymization 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.

2) Different Projection Probabilities: In this subsection, we quantify the conditions on p_a, p_u, l, and h when $p_a \neq p_u$ for structure based perfect de-anonymization 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 theorem 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 $\sigma \neq \sigma_0$. Note that, to avoid confusion, we consistently employ the same notations as in Theorems 1 and 2 in the remainder of this section.

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

Again, to guarantee the *uniqueness* of the perfect de-anonymization scheme σ_0 to cause the least DE when $p_a \neq p_u$, we quantify the conditions on p_a, p_u, l, and h as follows.

Theorem 4: 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 $2 \leq k \leq n$, it is a.a.s. *that there exists no de-anonymization scheme σ such that $\sigma \neq \sigma_0$ and $\Psi_\sigma \leq \Psi_{\sigma_0}$.*

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 $p_a \neq p_u$.

D. (1 − ϵ)-Perfect De-anonymization

In the aforementioned subsection, the conditions on perfect de-anonymization are quantified. Now, we study the conditions on $(1-\epsilon)$*-perfect de-anonymization.* Formally, we define a $(1-\epsilon)$*-perfect de-anonymization*, denoted by σ^ϵ, as a de-anonymization scheme under which at most $\epsilon|V^a| = \epsilon n$ users are tolerated to be incorrectly (unsuccessfully) de-anonymized, where $0 \leq \epsilon \leq 1$. Under the $(1-\epsilon)$-perfect de-anonymization 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 de-anonymization are quantified in Theorem 5. 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 de-anonymizaiton, our quantification takes σ_0 as the reference point. To make the paper more readable, we place all the proofs in the *Supplementary File.*

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 $k > \epsilon n$.

From Theorem 5, we can see that (i) for any de-anonymization 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 de-anonymization 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 tolerance of ϵn incorrect user mappings. As in the scenario of perfect de-anonymization, stronger conditions can be quantified to guarantee $(1-\epsilon)$-perfect de-anonymization schemes causing the least DE. The quantification is shown in Theorem 6, which can be proven by employing similar techniques as in Theorems 2 and 4. Therefore, we omit the detailed proof here. 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 de-anonymization scenario.

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 there exists no σ_k such that $k > \epsilon n$ and $\Psi_{\sigma_k} \leq \Psi_{\sigma_0}$.*

IV. Large-Scale Evaluation on Real World Datasets

According to our quantification, even without semantic/contextual priori knowledge, anonymized structural data can be de-anonymized perfectly or $(1-\epsilon)$-perfectly when certain structural conditions are satisfied. In this section, we conduct comprehensive evaluations of our de-anonymization quantification on 26 real world structural datasets.[8]

A. 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 [8], [29], [35]. Fortunately, it is not necessary to know the exact $p_{i,j}$ to numerically evaluate our de-anonymization 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., $p_{\mathbf{D}} \simeq \rho$ [8], [29]. On the other hand, we focus on demonstrating the statistical behavior of our perfect/$(1-\epsilon)$-perfect de-anonymization quantification. Therefore, we

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

TABLE I

DATA STATISTICS. n: NUMBER OF USERS; m: NUMBER OF EDGES; ρ: GRAPH DENSITY; $\overline{d}$: AVERAGE DEGREE; AND $p(i)$: FRACTION OF USERS WITH DEGREE LESS THAN OR EQUAL TO i

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

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 de-anonymization, (i) when $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{4\ln n+2}{2kn-k^2-k})$, $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \rightarrow 1$ *for any* $\sigma \neq \sigma_0$*; (ii) when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{2(k+3)\ln n+2}{2kn-k^2-k})$, *it is* a.a.s. *that there exists no* σ *such that* $\sigma \neq \sigma_0$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$.

Theorem 8: For $(1-\epsilon)$*-perfect de-anonymization, (i) when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{\ln n}{\epsilon n^2})$, $\Pr(\Psi_{\sigma_k} \geq \Psi_{\sigma_0}) \rightarrow 1$ *for any* σ_k *with* $k > \epsilon n$*; (ii) when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{\ln n}{n})$, *it is* a.a.s. *that there exists no* σ_k *such that* $k > \epsilon n$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$.

Now, based on Theorems 7 and 8, we evaluate our quantification on perfect and $(1-\epsilon)$-perfect de-anonymization.

B. Datasets

We evaluate our quantification on 26 datasets from multiple domains, including Social Network (SN) 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 [4], [33]–[35]. In Table I, 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 in the corresponding dataset.

Due to space limitations, we briefly introduce the datasets as follows. Detailed descriptions can be found in [4] and [33]–[35].

- **SN**. We employed 8 SN datasets in our evaluation as shown in Table I. Google+ is a SN developed by Google indicating the "circle" relationships (e.g., friends, families, colleagues) among people [33]. Twitter is a SN that enables users to send and read "tweets" [34]. LiveJournal is a SN that allows members to maintain journals, blogs, etc. [34]. Facebook is a SN where users are connected by "friendships" [34]. In the YouTube and Orkut SNs, users form "friendships" and create groups where other users can join [34]. Slashdot is a SN for sharing and maintaining technology-related news [34]. Pokec is also a "friendship" based SN [34].
- **LMSN**. Infocom consists of a Bluetooth contact trace and a coauthor network of Infocom 2006 conference attendees [4]. Smallblue consists of an *instant messenger* contact trace and a Facebook SN of the employees of a company [4]. Both Brightkite and Gowalla are consisting of a SN and a check-in trace of the SN users [34], [35].
- **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 [34]. DBLP is a collaboration network of researchers mainly in *Computer Science* [34].
- **Email and WikiTalk**. Enron and EuAll are two email communication networks [34]. WikiTalk is a network containing the discussion relationships among a group of users on Wikipedia [34].
- **AS**. AS733, Oregon, Caida, and Skitter are four AS graphs at different locations [34].
- **P2P**. Gnutella3, Gnutella4, and Gnutella5 are three P2P network graphs where nodes represent hosts in Gnutella and edges are connections between hosts [34].

Before evaluating our quantification, we preprocess the datasets as follows. First, we remove *isolated* users (or nodes) from a dataset if present (most of the datasets do not have isolated users). This is intuitively reasonable since we cannot leverage structural information to de-anonymize isolated users. Second, we do not consider the direction information of the directed data, i.e., all the datasets are represented by undirected graphs. This is because our network model is an undirected graph. Even direction takes some extra auxiliary information [3], we do not consider it in this paper and would include it in the future. More importantly, our quantification demonstrates that undirected structure information is powerful enough to de-anonymize structural data, which can also be seen in our following evaluation.

C. Evaluation on Perfect De-Anonymization Quantification

For each of the datasets considered, we represent it as graph G. Given $\wp$, G^a and G^u can be projected from G according two independent edge/relationship projection processes. Furthermore, the quantifications in Theorems 7 and 8 are

TABLE II

EVALUATION OF $(\Omega(f_{\mathbf{D}}), \Omega(n))$ IN PERFECT DE-ANONYMIZATION

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)

meaningful when n is a large number. Therefore, in the evaluation of perfect/$(1-\epsilon)$-perfect de-anonymization quantification, we also 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 de-anonymization under different projection probabilities $\wp$ are shown in Table II.

From Table II, we have the following observations.

- 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).
- 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 de-anonymization scheme *a.a.s.* inducing the least DE.
- 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 de-anonymization, which is consistent with our theoretical quantification. This comes from the similarity of their statistical $p_{\mathbf{D}}$. For perfect de-anonymization on datasets with different graph densities (with similar or different sizes), e.g., HepPh (n = 1.2E4, ρ = 1.87E-3) and 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 considered 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.
- From Table II, 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 $\wp \geq \Omega(.8)$, and Twitter is *a.a.s.* can be perfectly de-anonymized when $\wp \geq \Omega(.9)$. The perfect de-anonymization is due to their good structural characteristics, e.g., high average degree (from Table I, the average degree $\overline{d}$ is 76.3 for Orkut, 54.8 for Twitter, and 43.7 for Facebook), small percentage of nodes with a low degree ($p(1)$ is 2.2% for Orkut, 5.3% for Twitter, and 5.4% for Facebook).

D. Evaluation on (1 − ϵ)-Perfect De-Anonymization Quantification

Based on our quantification, the percentage of successfully de-anonymized users by any $(1-\epsilon)$-perfect de-anonymization 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 Table III.

From Table III, we make some important observations and comments as follows.

- When $\wp$ increases, more users can be de-anonymized for every dataset as expected. For example, when $\wp = .5$,

TABLE III
EVALUATION OF $\Omega(1-\epsilon)$ IN $(1-\epsilon)$-PERFECT DE-ANONYMIZATION

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%

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 Table III, similar de-anonymization phenomena applied to all the datasets, which is consistent with our quantification. The reason is straightforward. When $\wp$ increases, more edges/relationships appear in both G^a and G^u (the expected number of common edges is $m\wp^2$). Thus, the structural similarity between G^a and G^u is increased and more users can statistically be successfully de-anonymized with probability 1.

- Most of the existing structural datasets, including SN data, LMSN data, Email and Wiki data, AS data, P2P data, etc., 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 $\wp = .8$, Twitter can be completely de-anonymized when $\wp = .85$, and Google+ can be completely de-anonymized when $\wp = .95$. Even if a dataset cannot be completely de-anonymized, it may be partially de-anonymizable. 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.

- An interesting observation is that the de-anonymization results on two datasets with similar graph densities may be very different in practice. From Table II, for two datasets with similar graph densities, e.g., Google+ ($\rho = 8.24$E-6) and Skitter ($\rho = 7.73$E-6), the theoretical bounds on $(\Omega(f_{\mathbf{D}}), \Omega(n))$ for perfect de-anonymization are also similar. However, from Table III, the de-anonymization results of Google+ and Skitter are very different: when $\wp = .6$, the number of de-anonymizable users in Google+ (41.8%) is about twice of that 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 are only bounded by $\Omega(59.1\%)$. To study the reason for this fact, we need to consider the degree distribution of Google+ and Skitter in addition to the graph density (as well as $\Omega(f_{\mathbf{D}})$ and $\Omega(n)$). From Table I, the percentage of low degree users in Skitter ($p(1) = 12.8\%$ and $p(5) = 55.4\%$) is much higher than that in Google+ ($p(1) = 5.4\%$ and $p(5) = 27.3\%$). 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, 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 Tables I and III, if a dataset has a high average degree and a small percentage of low degree users, e.g., Orkut, Facebook, Twitter, Google+, 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, e.g., EuAll, Wiki, Caida, they are difficult to de-anonymize based solely on the structural information.

- Following the above observation, we find that there exists some difference between theory and practice on the dominating factor of de-anonymization. Theoretically, the graph density is a dominating factor on determining the bound

TABLE IV
EVALUATION OF $\Omega(n)$ IN $(1-\epsilon)$-PERFECT DE-ANONYMIZATION

Dataset	$\wp = .3$				$\wp = .6$				$\wp = .9$			
	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$
Google+	2.4E8	1.9E8	1.4E8	1.1E8	2.2E7	1.7E7	1.3E7	9.5E6	4.6E6	3.6E6	2.7E6	2.3E6
Twitter	1.4E7	1.1E7	8.4E6	6.2E6	1.2E6	9.6E5	7.3E5	5.4E5	2.5E5	2.3E5	2.3E5	2.3E5
LiveJournal	5.6E8	4.4E8	3.4E8	2.5E8	5.1E7	4.0E7	3.0E7	2.2E7	1.1E7	8.4E6	6.4E6	4.7E6
Facebook	1.1E5	9.0E4	6.9E4	5.2E4	9.2E3	7.2E3	5.5E3	4.1E3	2.0E3	2.0E3	2.0E3	2.0E3
YouTube	4.5E8	3.6E8	2.7E8	2.0E8	4.0E7	3.2E7	2.5E7	1.8E7	8.6E6	6.8E6	5.2E6	3.8E6
Orkut	7.5E7	5.9E7	4.6E7	3.5E7	6.7E6	5.3E6	4.1E6	3.1E6	1.5E6	1.5E6	1.5E6	1.5E6
Slashdot	9.7E6	7.7E6	5.9E6	4.4E6	8.5E5	6.7E5	5.2E5	3.8E5	1.7E5	1.4E5	1.1E5	7.7E4
Pokec	1.1E8	8.9E7	6.8E7	5.1E7	1.0E7	8.0E6	6.1E6	4.5E6	2.1E6	1.7E6	1.3E6	9.4E5
Infocom	1.5E4	1.3E4	1.1E4	9.0E3	1.2E3	9.8E2	7.8E2	6.8E2	2.5E2	2.5E2	1.7E2	1.7E2
Smallblue	2.2E4	1.8E4	1.5E4	1.2E4	1.8E3	1.4E3	1.2E3	8.8E2	3.4E2	3.2E2	2.2E2	2.2E2
Brightkite	1.3E7	1.0E7	7.7E6	5.7E6	1.1E6	8.8E5	6.7E5	4.9E5	2.3E5	1.8E5	1.4E5	1.0E5
Gowalla	3.6E7	2.9E7	2.2E7	1.6E7	3.2E6	2.5E6	1.9E6	1.4E6	6.7E5	5.3E5	4.0E5	3.0E5
HepPh	7.4E5	5.8E5	4.4E5	3.2E5	6.2E4	4.9E4	3.7E4	2.7E4	1.2E4	9.7E3	7.3E3	5.6E3
AstroPh	1.2E6	9.2E5	7.0E5	5.2E5	9.9E4	7.8E4	5.9E4	4.4E4	2.0E4	1.6E4	1.2E4	9.0E3
CondMat	3.9E6	3.1E6	2.4E6	1.9E6	3.4E5	2.7E5	2.1E5	1.6E5	6.9E4	5.5E4	4.2E4	3.2E4
DBLP	9.1E7	7.3E7	5.7E7	4.3E7	8.1E6	6.5E6	5.0E6	3.8E6	1.7E6	1.4E6	1.1E6	8.0E5
Enron	5.0E6	3.9E6	3.0E6	2.2E6	4.3E5	3.4E5	2.6E5	1.9E5	8.8E4	6.9E4	5.2E4	3.8E4
EuAll	1.5E8	1.2E8	9.3E7	7.0E7	1.3E7	1.1E7	8.3E6	6.2E6	2.8E6	2.2E6	1.7E6	1.3E6
Wiki	1.3E9	1.1E9	8.4E8	6.3E8	1.2E8	9.9E7	7.7E7	5.7E7	2.6E7	2.1E7	1.6E7	1.2E7
AS733	2.3E6	1.8E6	1.4E6	1.1E6	2.0E5	1.6E5	1.2E5	9.0E4	4.0E4	3.2E4	2.4E4	1.8E4
Oregon	3.1E6	2.5E6	1.9E6	1.4E6	2.7E5	2.1E5	1.6E5	1.2E5	5.5E4	4.3E4	3.3E4	2.4E4
Caida	1.1E7	8.9E6	6.9E6	5.1E6	9.8E5	7.8E5	6.0E5	4.5E5	2.0E5	1.6E5	1.2E5	9.1E4
Skitter	2.6E8	2.0E8	1.6E8	1.2E8	2.3E7	1.8E7	1.4E7	1.0E7	4.9E6	3.9E6	3.0E6	2.2E6
Gnutella3	9.0E6	7.1E6	5.5E6	4.0E6	7.8E5	6.2E5	4.8E5	3.5E5	1.6E5	1.3E5	9.7E4	7.1E4
Gnutella4	1.3E7	1.0E7	8.0E6	5.9E6	1.1E6	9.0E5	6.9E5	5.1E5	2.3E5	1.9E5	1.4E5	1.0E5
Gnutella5	2.3E7	1.9E7	1.4E7	1.1E7	2.1E6	1.6E6	1.3E6	9.3E5	4.3E5	3.4E5	2.6E5	1.9E5

of $(\Omega(f_{\mathbf{D}}), \Omega(n))$ (Table II). In practice, the degree distribution and the average degree have more impact on the de-anonymization results (Table III). 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 de-anonymization conditions are satisfied.

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

- When ϵ is fixed, the impact of $\wp$ on $\Omega(n)$ in $(1-\epsilon)$-perfect de-anonymization is similar to that in perfect de-anonymization, 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 $\sigma_{k'}$ ($k' > \epsilon n$).
- When $\wp$ is fixed, $\Omega(n)$ is also decreasing as ϵ increases. For instance, when $\wp = 0.6$, $\Omega(n)$ is decreased from 2.2E7 to 9.5E6 for Google$+$ when ϵ is increased from .1 to .4. This is because when ϵ increases, more DE is tolerated, and thus a 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.
- As in the perfect de-anonymization scenario, graph density is an important factor that impacts $\Omega(n)$. 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.

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

- Theoretically, the condition on the lower bound of $\wp$ is very loose, e.g., when $\epsilon = .1$, $\Omega(\wp) = 1.1$E-7 for Google$+$ and $\Omega(\wp) = 1.7$E-7 for Orkut, which suggests that $(1-\epsilon)$-perfect de-anonymization 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 $\epsilon = .1$, $\Omega(n) = 2.2$E28 for Google$+$ and $\Omega(n) = 2.0$E27 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 Tables II, III, and IV).
- From Table V, we can see that the conditions on $\Omega(f_{\mathbf{D}})$ and $\Omega(n)$ exhibit the same behavior as in perfect de-anonymization, 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.
- From Table V, 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 de-anonymization scenario.

V. OPTIMIZATION BASED DE-ANONYMIZATION PRACTICE

In Section III, we comprehensively quantify conditions for perfect de-anonymization and $(1-\epsilon)$-perfect de-anonymization. Based on our large-scale study of 26 real world datasets in Section IV, we find most, if not all, existing

TABLE V

EVALUATION OF $(\Omega(\wp), \Omega(f_{\mathbf{D}}), \Omega(n))$ IN $(1-\epsilon)$-PERFECT DE-ANONYMIZATION

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)

structural datasets are de-anonymizable partially or completely (Table III). Interestingly, our de-anonymization quantification naturally leads to a de-anonymization scheme, denoted by $\mathfrak{A}^*$. Basically, $\mathfrak{A}^*$ can be implemented as follows: we can calculate the DE caused by each σ_k $(1 \leq k \leq n!)$ 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 de-anonymization 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.

A. Optimization Based De-Anonymization

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 β is a user-input parameter (a non-negative integer) and d_k^i $(1 \leq k \leq \beta)$ 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 $d_{|N_i^a|+1}^i = d_{|N_i^a|+2}^i = \cdots = d_\beta^i = \Delta^a$ (resp., $d_{|N_i^u|+1}^i = d_{|N_i^u|+2}^i = \cdots = d_\beta^i = \Delta^u$), 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 $(1 \leq k \leq K)$ 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). If there is a tie, we randomly pick one reference user from the users with the same degree. 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 de-anonymization scheme σ with $\sigma(v_k) = u_k$ $(1 \leq k \leq L)$. Intuitively, V_L^a and U_L^u can be used as auxiliary information for future de-anonymization. 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 $(1 \leq k \leq L)$ 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 $f_c(i) = \sum\limits_{j \in S^a \setminus \{i\}} \frac{1}{h(i,j)}$ (resp., $f_c(i) = \sum\limits_{j \in S^u \setminus \{i\}} \frac{1}{h(i,j)}$), 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* of a user, we introduce a sampling closeness centrality feature, which can characterize the global

Algorithm 1 Optimization Based De-Anonymization (ODA)

```
1 Define $\Lambda^a = \Lambda^u = \emptyset$;
2 while true do
3     $\Lambda^a$ = GetTopDegree($V^a, \alpha$), $\Lambda^u$ = GetTopDegree($V^u, \alpha$);
4     for every $i \in \Lambda^a$, compute a candidate mapping set $\mathcal{C}(i)$ = GetTopSimilarity($i, \Lambda^u, \gamma$);
5     apply the consistent rule and pruning rule to find the de-anonymization scheme $\sigma(\Lambda^a) \in \prod_{i \in \Lambda^a} (i \times \mathcal{C}(i))$ which induces the least DE $\Psi_{\sigma(\Lambda^a)}$, denoted by $\sigma^*(\Lambda^a) = \{(i_1, j_1), (i_2, j_2), \cdots, (i_\alpha, j_\alpha)\}$;
6     for each $(i, j) \in \sigma^*(\Lambda^a)$, if $\phi(i, j) \geq \theta$ then
7         accept the mapping $(i, j)$;
8         $V^a = V^a \setminus \{i\}$, $V^u = V^u \setminus \{j\}$;
9     if no mapping in $\sigma^*(\Lambda^a)$ is accepted, break;
```

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 III, $\mathfrak{A}^*$ is inherently an optimization based algorithm with the objective of minimizing the DE Ψ_{σ_k}, which is different from most of existing de-anonymization algorithms (heuristics based) [2]–[4]. 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 de-anonymization 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 $\psi_{i,j} = |f_d(i) - f_d(j)| + (1 - \phi(i,j)) \cdot |f_d(i) - f_d(j)|$[9] 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 de-anonymization 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 $i \in \Lambda^a$, 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 ($\phi(i, \cdot)$) 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-de-anonymization scheme* $\sigma^*(\Lambda^a)$ which induces the least DE. We explain the idea of ODA in detail as follows. In Line 3, we initialize the target de-anonymization set Λ^a and the candidate mapping set Λ^u. From the initialization, $|\Lambda^a|, |\Lambda^u| \leq \alpha$ (since it is possible $|V^a|, |V^u| \leq \alpha$), where 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 de-anonymization 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 de-anonymization scheme $\sigma(\Lambda^a)$ consistent, i.e., no *mapping conflict* 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 de-anonymization scheme should be avoided. Note that, it 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 de-anonymization schemes whose DE is larger than the current known least DE. For instance, let $\sigma^*(\Lambda^a)$ be the de-anonymization scheme having the least DE after testing k possible de-anonymization schemes. Then, when testing the $(k+1)$-th possible de-anonymization 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 de-anonymization 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, the solution of $\mathfrak{A}^*$ is the optimum solution). This is because we try to seek a tradeoff between computational feasibility and de-anonymization 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 that had been rejected, they will be re-considered in the following iterations for possible better de-anonymizations. If no mapping can be accepted, we stop ODA. Subsequently, we analyze the time and space complexities of ODA in the following theorem. The proof is placed in the *Supplementary File* for readability.

[9] In the definition, $|f_d(i) - f_d(j)|$ measures the *absolute* neighborhood difference between i and j under any de-anonymization scheme. Further, $\phi(i,j)$ measures the structural similarity between i and j. Then, a smaller $\psi_{i,j}$ (the DE to map i to j) is induced when i and j are more structurally similar; otherwise, a larger $\psi_{i,j}$.

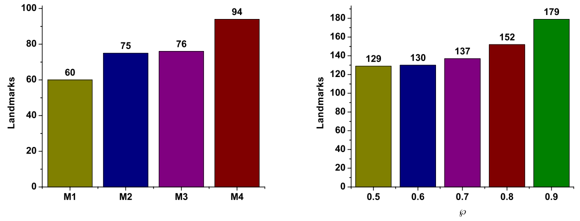
Theorem 9: (i) The space complexity of ODA is $O(\min\{n^2, m+n\})$. *(ii) Let* γ *be some constant value,* $\alpha = \Theta(\log n)$, *and* Γ *be the average number of accepted mappings in each iteration of ODA. Then, the time complexity of ODA is* $O(m + n\log n + n^{\Theta(1)\log\gamma+1}/\Gamma)$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

Finally, we make some remarks on ODA as follows.

- ODA is a *cold start* algorithm, i.e., we do not need any priori knowledge, e.g., the seed mapping information [2]–[4], to bootstrap the de-anonymization process. Furthermore, unlike existing de-anonymization algorithms [2]–[4] which consist of two phases (*landmark/seed identification phase* and *de-anonymization propagation phase*), ODA is a single-phase algorithm. Interestingly, ODA itself can act as a *landmark identification algorithm*. From our experiment (Section V-B), ODA can de-anonymize the 60-180 Top-degree users in Gowalla and Google+ (see Table I) perfectly, which can serve as landmarks (V_L^a and U_L^u) for future de-anonymization. In addition, ODA as a landmark identification algorithm is much faster than that in [3] (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 [4] (with complexity of $k!$, could be computationally infeasible for a PC when $k \geq 20$).
- Similar to $\mathfrak{A}^*$, ODA is an optimization based de-anonymization scheme, which is different from most of existing heuristics based solutions [2]–[4]. 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 de-anonymization scheme*.
- In ODA, we seek an adjustable tradeoff between de-anonymization 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 preserving satisfiable de-anonymization performance.
- ODA is a general framework. Line 5 can also be implemented by seeking a *maximum weighted bipartite graph matching* on a *weighted bipartite graph* $G(\Lambda^a \cup \Lambda^u, \bigcup\limits_{i\in\Lambda^a}(i \times \mathcal{C}(i)))$, where the weight on each edge is $\phi(i,j)$ $(i \in \Lambda^a, j \in \mathcal{C}(i))$.
- In practice, it is possible that V^a and V^u are not generated by the exactly same group of users. 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 de-anonymization). One better solution could be estimating the overlap between G^a and G^u first using the technique in [5], and then applying ODA to the overlap to achieve better performance.

B. Experimental Evaluation and Analysis

1) Datasets and Setup: We evaluate the performance of ODA on two real world datasets: Gowalla and Google+ (see the basic information in Section IV). Gowalla is a location based social network and consists of two different datasets [34], [35]. The first dataset is a spatiotemporal mobility trace consisting of 6,442,890 *check-ins* generated by 196,591 users. Each check-in has the format of <UserID, latitude, longitude, timestamp, location ID>. The second dataset is a social graph (950,327 edges) of the same 196,591 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 [35] on the mobility trace to construct four graphs with different *recalls* and *precisions*, denoted by $M1, M2, M3$, and $M4$, respectively (recall $= \frac{\text{true positive}}{\text{true positive+false negative}}$ and precision $= \frac{\text{true positive}}{\text{true positive+false positive}}$). 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 considering dataset is the Google+ dataset in Section IV, which has 4,692,671 users and 90,751,480 edges. Given some projection probability $\wp \in [0.5, 0.9]$, We first use the *projection process* in Section III 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 contextual domain (social data) with the anonymized data (mobility trace) in Gowalla* while the auxiliary and anonymized data are from the same domain in Google+.



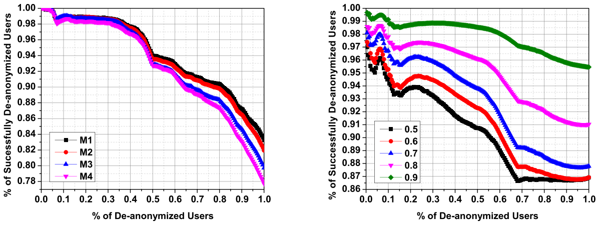
(a) (b)

Fig. 2. Landmark identification. $c_1, c_2 \in [0.1, 0.3], c_3 \in [0.4, 0.8]$, $c_4 = 0, \alpha \in [10, 30], \gamma \in [1, 4]$. (a) Gowalla. (b) Google+.

All the experiments are implemented on a PC with 64 bit Ubuntu 12.04 LTS operating system, Intel Xeon E5620 CPU (2.4GHz × 8 Threads), 48GB memory, and 2 disks with 8TB storage. When de-anonymizing Google+, each experiment is repeated five times (since G^a and G^u are randomly generated) and the results are the average values of these five runs. Here, we only show the de-anonymization results. More experiments/analysis on *de-anonymization error* and *time consumption* can be found in the *Supplementary File*.

2) Results:

a) 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 de-anonymization 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



(a) (b)

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

$\wp = 0.7$, the first 137 users in Google+ are perfectly de-anonymizable. According to ODA, the identified landmarks can serve as references for future de-anonymization.

From Fig. 2 (a), we can see that 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. Similarly, 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.

b) De-Anonymization 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 de-anonymization rate is higher for large-degree users than that of 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 de-anonymization performance is impacted more by precision. Generally, high precision implies that this dataset is more de-anonymizable, e.g. $M4$. This is because high precision implies a low false positive, which can be viewed as *noise* in practice, and thus the de-anonymization 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 de-anonymization performance of ODA on Gowalla and Google+ is better than the evaluation results shown in Table III, e.g., when $\wp = 0.9$, Table III 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 Table III 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 in different scenarios. Thus, structure based de-anonymization is powerful in practice.

VI. Conclusion and Future Work

In this paper, we study the quantification, practice, and implications of structural data de-anonymization. First, for the first time, we address several fundamental open problems in the data de-anonymization research by quantifying the conditions for *perfect de-anonymization and* $(1 - \epsilon)$-*perfect de-anonymization under a general data model*. This remedies the gap between structural data de-anonymization 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. We also quantitatively demonstrate the necessary conditions and reasons for the de-anonymizability of the 26 datasets. Third, following our quantification, we propose a practical de-anonymization technique that is a *cold start single-phase Optimization based De-Anonymization* (ODA) algorithm. We also analyze ODA theoretically and experimentally. The experimental results show that $77.7\% - 83.3\%$ of the users in Gowalla (196,591 users, 950, 327 edges) and $86.9\% - 95.5\%$ of the users in Google+ (4,692,671 users, 90,751,480 edges) can be de-anonymized, which implies structure based de-anonymization is implementable and powerful in practice. Finally, we conclude some implications from our findings.

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 ODA to make it more efficient and robust; (ii) Since existing anonymization techniques are vulnerable to structure based de-anonymization attacks, we propose to develop application based effective schemes against such attacks; (iii) In our quantification, we assume $V^a = V^u$. We plan to remove this assumption by quantifying the de-anonymizability of structural data when $V^a \neq V^u$; (iv) 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 (v) 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 de-anonymization attacks.

Acknowledgment

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.

References

[1] S. Ji, W. Li, M. Srivatsa, and R. Beyah, "Structural data de-anonymization: Quantification, practice, and implications," in *Proc. ACM CCS*, 2014, pp. 1040–1053.

[2] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou r3579x? Anonymized social networks, hidden patterns, and structural steganography," in *Proc. 16th Int. Conf. WWW*, 2007, pp. 181–190.

[3] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proc. 30th IEEE Symp. SP*, May 2009, pp. 173–187.

[4] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social network as a side-channel," in *Proc. ACM Conf. CCS*, 2012, pp. 628–637.

[5] S. Ji, W. Li, M. Srivatsa, J. S. He, and R. Beyah, "Structure based data de-anonymization of social networks and mobility traces," in *Proc. 17th Int. Conf. ISC*, 2014, pp. 237–254.

[6] S. Ji, W. Li, P. Mittal, X. Hu, and R. Beyah, "SecGraph: A uniform and open-source evaluation system for graph data anonymization and de-anonymization," in *Proc. 24th USENIX Secur. Symp.*, 2015, pp. 303–318.

[7] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel, "A practical attack to de-anonymize social network users," in *Proc. IEEE Symp. SP*, May 2010, pp. 223–238.

[8] P. Pedarsani and M. Grossglauser, "On the privacy of anonymized networks," in *Proc. 17th ACM SIGKDD Int. Conf. KDD*, 2011, pp. 1235–1243.

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

[10] K. Liu and E. Terzi, "Towards identity anonymization on graphs," in *Proc. SIGMOD*, 2008, pp. 93–106.

[11] N. Li, W. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy," in *Proc. 7th ASIACCS*, 2012, pp. 32–43.

[12] C. Dwork, "Differential privacy," in *Proc. 33rd ICALP*, 2006, pp. 1–12.

[13] A. Korolova, R. Motwani, S. U. Nabar, and Y. Xu, "Link privacy in social networks," in *Proc. 17th ACM CIKM*, 2008, pp. 289–298.

[14] 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," in *Proc. WWW*, 2009, pp. 531–540.

[15] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall, "802.11 user fingerprinting," in *Proc. MobiCom*, 2007, pp. 99–110.

[16] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in *Proc. WWW*, 2010, pp. 61–70.

[17] S. Han *et al.*, "Expressive privacy control with pseudonyms," in *Proc. SIGCOMM*, 2013, pp. 291–302.

[18] P. Mittal, M. Wright, and N. Borisov, "Pisces: Anonymous communication using social networks," in *Proc. NDSS Symp.*, 2013, pp. 1–18.

[19] J. Kannan, G. Altekar, P. Maniatis, and B.-G. Chun, "Making programs forget: Enforcing lifetime for sensitive data," in *Proc. 13th USENIX Conf. HotOS*, 2013, p. 23.

[20] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "COMPA: Detecting compromised accounts on social networks," in *Proc. NDSS Symp.*, 2013, pp. 1–17.

[21] K. Singh, S. Bhola, and W. Lee, "xBook: Redesigning privacy control in social networking platforms," in *Proc. 18th Conf. USENIX Secur. Symp.*, 2009, pp. 249–266.

[22] P. Hornyack, S. Han, J. Jung, S. Schechter, and D. Wetherall, "These aren't the droids you're looking for: Retrofitting android to protect data from imperious applications," in *Proc. 18th ACM Conf. CCS*, 2011, pp. 639–652.

[23] M. Egele, C. Kruegel, E. Kirda, and G. Vigna, "PiOS: Detecting privacy leaks in iOS applications," in *Proc. NDSS Symp.*, 2011, pp. 1–15.

[24] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, "SybilLimit: A near-optimal social network defense against sybil attacks," in *Proc. IEEE Symp. SP*, May 2008, pp. 3–17.

[25] H. Yu, C. Shi, M. Kaminsky, P. B. Gibbons, and F. Xiao, "DSybil: Optimal sybil-resistance for recommendation systems," in *Proc. 30th IEEE Symp. SP*, May 2009, pp. 283–298.

[26] L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi, "SoK: The evolution of sybil defense via social networks," in *Proc. IEEE Symp. SP*, May 2013, pp. 382–396.

[27] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Proc. IEEE Symp. SP*, May 2011, pp. 247–262.

[28] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec, "Protecting location privacy: Optimal strategy against localization attacks," in *Proc. ACM Conf. CCS*, 2012, pp. 617–627.

[29] M. E. J. Newman, *Networks: An Introduction*. London, U.K.: Oxford Univ. Press, 2010.

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

[31] B. Bollobás, *Random Graphs*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[32] J. Riordan, *An Introduction to Combinatorial Analysis*. New York, NY, USA: Wiley, 1958.

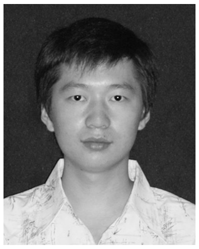
[33] N. Z. Gong *et al.*, "Evolution of social-attribute networks: Measurements, modeling, and implications using Google+," in *Proc. ACM Conf. IMC*, 2012, pp. 131–144.

[34] J. Leskovec and A. Krevl. (Jun. 2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. [Online]. Available: http://snap.stanford.edu/data.

[35] H. Pham, C. Shahabi, and Y. Liu, "EBM: An entropy-based model to infer social strength from spatiotemporal data," in *Proc. SIGMOD*, 2013, pp. 265–276.

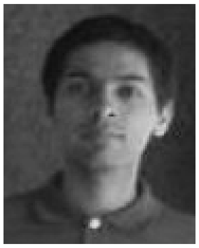
[36] C. Shah, R. Capra, and P. Hansen, "Collaborative information seeking," *Computer*, vol. 47, no. 3, pp. 22–25, 2014.

[37] Z. Xu, J. Ramanathan, and R. Ramnath, "Identifying knowledge brokers and their role in enterprise research through social media," *Computer*, vol. 47, no. 3, pp. 26–31, Mar. 2014.



Shouling Ji received the B.S. (Hons.) and M.S. degrees in computer science from Heilongjiang University, the Ph.D. degree in computer science from Georgia State University, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology. He is currently a ZJU 100-Young Professor with the College of Computer Science and Technology, Zhejiang University, and a Research Faculty Member with the School of Electrical and Computer Engineering, Georgia Institute of Technology. His current research interests include big data security and privacy, password security, and wireless networks. He is a member of IEEE and ACM and was the Membership Chair of the IEEE Student Branch at Georgia State University (2012–2013).

Weiqing Li received the B.S. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, where he is currently pursuing the M.S. degree. His research interests include big data privacy and network security. He is a student member of ACM and IEEE.



Mudhakar Srivatsa has been a Research Scientist with the Network Technologies Department, IBM Thomas J. Watson Research Center, since 2007. He is currently a Research Manager of the Mobile Network Analytics Team, an IBM Master Inventor, and an IEEE Senior Member. His expertise is in network analytics and secure information flow.

Raheem Beyah received the B.Sc. degree in electrical engineering from North Carolina A&T State University in 1998, and the master's and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology (Georgia Tech) in 1999 and 2003, respectively. He was an Assistant Professor with the Department of Computer Science, Georgia State University, a Research Faculty Member with the Communications Systems Center (CSC), Georgia Tech, and a Consultant with the Network Solutions Group, Accenture. He is currently an Associate Professor with the School of Electrical and Computer Engineering, Georgia Tech, where he leads the Communications Assurance and Performance Group and is a member of CSC. His research interests include network security, wireless networks, network traffic characterization and performance, and critical infrastructure security. He is a member of AAAS and ASEE, a Lifetime Member of NSBE, and a Senior Member of ACM and IEEE. He received the National Science Foundation CAREER Award in 2009, and was selected for DARPA's Computer Science Study Panel in 2010.