On Your Social Network De-anonymizablity: Quantification and Large Scale Evaluation with Seed Knowledge

Shouling Ji*, Weiqing Li*, Neil Zhenqiang Gong†, Prateek Mittal‡ and Raheem Beyah*
*Georgia Institute of Technology
Email: {sji, wli64}@gatech.edu, rbeyah@ece.gatech.edu
†University of California, Berkeley
Email: neilz.gong@berkeley.edu
‡Princeton University
Email: pmittal@princeton.edu

***Abstract*—In this paper, we conduct the first comprehensive quantification on the *perfect de-anonymizability* and *partial de-anonymizability* of real world social networks with seed information in general scenarios, where a social network can follow an arbitrary distribution model. This quantification provides the theoretical foundation for existing structure based de-anonymization attacks (e.g., [1][2][3]) and closes the gap between de-anonymization practice and theory. Besides that, our quantification can serve as a *testing-stone* for the effectiveness of anonymization techniques, i.e., researchers can employ our quantified structural conditions to evaluate the potential de-anonymizability of the anonymized social networks. Based on our quantification, we conduct a large scale evaluation on the de-anonymizability of 24 various real world social networks by quantitatively showing: 1) the conditions for perfect and $(1-\epsilon)$ de-anonymization of a social network, where ϵ specifies the *tolerated de-anonymization error*, and 2) the number of users of a social network that can be successfully de-anonymized. Furthermore, we show that, both theoretically and experimentally, the overall structural information based de-anonymization attack is much more powerful than the seed knowledge-only based de-anonymization attack, and even without any seed information, a social network can be perfectly or partially de-anonymized. Finally, we discuss the implications of this work. Our findings are expected to shed light on the future research in the structural data anonymization and de-anonymization area, and to help data owners evaluate their structural data vulnerability before data sharing and publishing.**

I. Introduction

Fueled by the rapid advancements in information and mobile computing technologies, social networks have become deeply integrated in people's lives. Further, social networks produce a significant amount of social data that contains their users' detailed personal information. [2][3][19]. To protect users' privacy, data owners (e.g., companies, government, hospitals) usually anonymize their data before it is shared, transferred, and/or published. Generally, the data anonymization techniques can be characterized into three classes: *naive ID removal*, *k-anonymization (including randomly adding/deleting edges)* [11][12], and *differential privacy* [13][14]. The naive ID removal method has been proven extremely vulnerable to state-of-the-art structure based de-anonymization attacks [2][3]. For k-anonymization, it also cannot be employed to defend against structure based de-aonymization attacks for real world social networks due to its limitations, e.g., it is not scalable (an NP-hard problem [16]), richer information is available to adversaries. Differential privacy (and its variants) is initially designed to protect the privacy of data in an *interactive query* [13][14]. However, the structure based de-anonymization attack is actually a *non-interactive query* from the perspective of database interaction. Recent efforts have proposed using *differentially private graph models* [15]. However, the *differential privacy parameter* still has to be determined before data sharing, which makes this method ineffective in defending against structure based de-anonymization attacks [2][3]. Furthermore, the richer auxiliary information (e.g., seeds information) available to adversaries, which is difficult to control, makes the above anonymization technique more vulnerable. Consequently, *data anonymization is a challenging and open problem* (see more discussion in Section II).

Due to the vulnerability of existing anonymization schemes, it has been experimentally demonstrated that the the emerging *structure based de-anonymization attacks* can break the privacy of social networks effectively based only on the data's structural information, e.g., Narayanan and Shmatikov's de-anonymization attack [2], Srivatsa and Hicks' de-anonymization attack [3]. Although the de-anonymizability of social networks has been shown by experimental results (heuristic algorithms) in [2][3], it is still important to understand *why social networks are vulnerable to structure based de-anonymization attacks? how de-anonymizable a social network is? and how many users within a social network can be successfully de-anonymized?* Currently, some preliminary analysis was conducted on the de-anonymizability of social networks under the the *Erdös-Rényi* (ER) random graph model, the *preferential attachment* model, or the *configuration model*

NDSS '15, 8-11 February 2015, San Diego, CA, USA
Copyright 2015 Internet Society, ISBN 1-891562-38-X
http://dx.doi.org/10.14722/ndss.2015.23096

[6][7][8][9]. On one hand, these existing works open the door to research on quantifying the de-anonymizability of social networks. On the other hand, all the existing works have several limitations, e.g., most do not consider seed information or overlook other more powerful structural information (e.g., the structural information among non-seed users), they assume an unrealistic network model (e.g., the ER model), and/or they make unrealistic assumptions (e.g., having a regime of dense seeds available). These limitations prevent existing analysis from being applicable to real world social networks (see more discussion in Section II).

Contributions: Aiming at addressing these open problems, we study the de-anonymizability of social networks. Specifically, our contributions can be summarized as follows.

(i) To the best of our knowledge, we conduct the first theoretical quantification on the *perfect* and *partial de-anonymizability* of social networks in general scenarios, where the social network can follow an arbitrary network model. Therefore, our quantification can be applied to real world social networks to evaluate their perfect or partial de-anonymizability, i.e., our quantification can quantitatively demonstrate the vulnerability of real world social networks to existing structure based de-anonymization attacks (e.g., [1][2][3]). More importantly, our quantification provides the theoretical foundation for existing structure based de-anonymization attacks (e.g., [1][2][3]), which closes the gap between practice and theory. Besides that, since our quantification specifies the structural conditions between an anonymized social network and an auxiliary social network for perfect and partial de-anonymizability, our quantification can serve as a *testing-stone* for the effectiveness of anonymization techniques, i.e., researchers can employ our quantified structural conditions to evaluate the potential de-anonymizability of the anonymized social networks.

(ii) Based on our quantification, we implement the first large scale evaluation of the perfect and partial de-anonymizability of 24 various real world social networks. In our evaluation, we show the conditions of perfectly and partially de-anonymizing a social network; how de-anonymizable a social network is according to its topological properties; and how many users of a social network can be successfully de-anonymized. Our evaluation results demonstrate that most social networks, if not all, can be perfectly or at least partially de-anonymized depending on their structural properties.

(iii) Based on our results, we find that compared to the structural information associated with known seed users, the other structural information (the structural information among anonymized users) is also useful in improving structure based de-anonymization attacks. We show that, both theoretically and experimentally, the overall structural information based de-anonymization is more powerful, and a social network is perfectly or partially de-anonymizable even without any seed information. As a result, this finding provides the foundation of the implication that *unlike existing seed based de-anonymization techniques [1][2][3], one can design new effective de-anonymization attacks without seed information*.

(iv) We discuss the implications and future work of this paper. Our quantification and evaluation enable various s-takeholders (e.g., researchers and practitioners) to understand the theoretical foundation of structure based de-anonymization attacks and their effectiveness in attacking various real world social networks (in other words, the vulnerability of real world social networks). Therefore, our work can shed light on the future research of the structural data anonymization and de-anonymization area, and encourage data owners to develop better privacy protection policies.

II. Related Work

Due to space limitations, we only discuss the most closely related work. Readers can find more related work on de-anonymization in [2][3], on de-anonymization quantification in [7][8][9], on anonymization techniques in [11][12], and on privacy preservation schemes in [13][14].

A. Anonymization

To protect users' privacy, several anonymization techniques have been developed which can be placed into three classes: *naive ID removal*, *k-anonymization* [11][12], and *differential privacy* [13][14]. The *naive ID removal* scheme has proven to be vulnerable to structure based de-anonymization attacks [1][2][3]. However, naive ID removal is still the most widely used method to anonymize data for publishing/sharing/transferring [2][3]. This is probably because of two reasons. First, it is the simplest manner to "anonymize" data and it is scalable so it can be applied to large scale datasets in a straightforward manner. Second, the lack of education on the vulnerability of this method leads data owners to choose this anonymization scheme.

k-anonymization is a sophisticated anonymization technique, under which each user cannot be distinguished with at least $k-1$ other individuals with respect to their local structure [11][12]. However, several limitations make k-anonymization fail to defend against the newly designed structure based de-anonymization attacks (e.g., [2][3]). First, k-aonymization is not computationally scalable, which implies it cannot be extended to large scale social networks. Second, k-anonymization relies on data's syntactic property. Even if the k-anonymity is satisfied, it is still inefficient against the state-of-the-art de-anonymization attacks, which are based only on structural information [2]. Finally, the adversary may have much richer information than assumed in the k-anonymization. For instance, the known seed information together with the data's structure are sufficient to perfectly or at least partially de-anonymize a social network according to our quantification.

Differential privacy (and its variants) aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the possibility of leaking privacy [13][14], i.e., differential privacy is designed to protect data's privacy in *interactive queries* [13][14]. However, the structure based de-anonymization attacks we studied in this paper are *non-interactive queries* from the perspective of databases. There has also been proposed techniques to share graphs using *differentially private graph models* [15]. However, the *differential privacy parameter* still has to be determined before data sharing, which makes this method ineffective in defending against structure based de-anonymization attacks [2][3].

B. Structure based De-anonymization

Structure based de-anonymization was introduced in [1], where Backstrom et al. introduced both active attacks and passive attacks to de-anonymize social data. For the active attack, the adversary should create a number of "sybil" nodes and build relationships between sybil nodes and target nodes before data release. Several reasons limit the practicality of active attacks. A direct limitation is that the active attack is not scalable and difficult to control because the amount of social data continues to increase [23]. Furthermore, sybil defense schemes [24] make this attack even more difficult. On the other hand, in real world social networks, target nodes have no reason to respond to the connection requests from strange sybil nodes. For the passive attack in [1], the adversary can breach the privacy of users with whom they are linked, which is again suitable for small social networks and difficult to extend to large scale social datasets. In [2], Narayanan and Shmatikov extended the de-anonymization attack to large-scale directed social networks, i.e., the social data carries direction information which can be used as auxiliary knowledge. They designed a de-anonymization algorithm with two phases: *seed identification* and *propagation*. In [3], Srivatsa and Hicks presented the first de-anonymization attack on mobility traces while using social networks as a side-channel. However, scalability is a significant limitation of the algorithms in [3]. In [4], Ji et al. designed an Adaptive De-Anonymization (ADA) framework for the scenario that the anonymized and auxiliary graphs have partial overlap. ADA also consists of a seed identification phase and a propagation phase. Recently, Nilizadeh et al. proposed a *community-enhanced* de-anonymization scheme of social networks [5]. Under this scheme, *community-level de-anonymization* is first conducted. Subsequently, the obtained information is employed to enhance the *user-level de-anonymization*.

In summary, all the mentioned de-anonymization algorithms are heuristic. None of them study the de-anonymizability of structural data (including social networks, mobility traces, etc.) quantitatively or theoretically. This motivates us to quantify the perfect and partial de-anonymizability of social networks, which closes the gap between existing structure based de-anonymization practice and its theoretical foundation.

There are also other kinds of de-anonymization attacks on social networks. For example, in [10], Wondracek et al. designed a de-anonymization attack on social networks based on the *group membership* information. To implement this attack, the adversary should collect necessary group membership information (which is semantic information) by "history stealing" on browsers.

C. Quantification of Social Network De-anonymizablity

Recently, the problem of understanding why social networks can be de-anonymized based on structural information has garnered a lot of attention [6][7][8][9]. In [6], Pedarsani and Grossglauser studied the de-anonymizability of social networks under the *ER random graph* model. When using an ER model to describe social networks, the primary limitation is that the degree distribution of social networks should follow the *Poisson distribution*. However, the degree distribution of real world social networks may follow any distribution (e.g., power-law distribution) [19]. Furthermore, it is seldom to see, if not impossible to see, the degree distribution of any social network following the Poisson distribution [19]. For example, none of the 24 real world social networks considered in this paper follow the Poisson distribution. Consequently, the quantification under the ER model is mathematically meaningful, however, cannot be applied to real world social networks. Nevertheless, the ER model is a nice mathematical tool to simplify the theoretical quantification, and thus the quantification under the ER model can shed light on the quantification in general scenarios. Another drawback of [6] is that it did not consider seed information.

In [7], Ji et al. studied the de-anonymizability of social networks under the *configuration model*. They also proposed a practical optimization-based de-anonymization algorithm. However, in their quantification and analysis, seed information is not considered. In addition, compared to the employed configuration model, our results are more general since we do not make assumption on the data distribution. Finally, the quantification in [7] can be viewed as a special case of our results in Section V (the $\Lambda = 0$ case). Therefore, our results in this paper are more general and practical.

In [8], Yartseva and Grossglauser studied the performance of percolation graph matching, which is a correlated problem of the de-anonymizablity of social networks. However, the analysis derived in [8] also has some limitations. First, the analysis in [8] is based on the ER model, which makes it impractical. Second, the analysis in [8] only considers seed information. According to our quantification and evaluation, we found that the other structural information (i.e., the structural information among anonymized users) is more powerful in improving the performance of de-anonymization attacks. We also demonstrate in this paper that a social network can be perfectly or partially de-anonymized even without any seed information. Third, the analysis in [8] only considers users with degree no less than 4. However, many real world social networks may have a large amount of users with degree less than 4, e.g., from our statistics in Section VI, 17.1% Google+ users, 40.03% LiveJournal users, and 77.33% YouTube users have degree less than 4.

In [9], Korula and Lattanzi studied the reconciliation problem for social networks, which is another correlated problem of the de-anonymizability of social networks (the reconciliation-ability of social networks is equivalent to their de-anonymizability). The analysis in [9] is conducted under the ER model and the Preferential Attachment model. Again, in [9], only seed information is considered in quantifying the de-anonymizability of social networks. As we pointed out before, in this paper, we find that the other structural information is more powerful in a de-anonymization attack, i.e., even without any seed knowledge, social networks can be perfectly or partially de-anonymized only based on the structural information. Another important limitation is that the quantification in [9] is valid under the assumption of having a regime of dense seeds available. However, this assumption is usually not true for large scale real world social networks.

In summary, the following aspects distinguish our work from existing works. First, for the first time, we quantified the perfect and partial de-anonymizability of social networks with seed information consideration under an arbitrary model. Second, compared to seed information, we find that the other structural information is more powerful in de-anonymization attacks. We demonstrate that, both theoretically and experimentally, even without any seed information, most real world social networks can also be de-anonymized perfectly or partially only based on the structural information. Third, we conduct the first large scale evaluation on the perfect and partial de-anonymizability of 24 various real world social networks. We quantitatively demonstrate the conditions on de-anonymizing real world social networks and how many users can be successfully de-anonymized.

III. System Model, Assumption, and Definition

In this paper, given some seed knowledge, we quantify the de-anonymizability of social networks based on their structural information. It is natural to model social networks as graphs where nodes represent users and edges represent social ties (friends, contacts, etc.) among the users [1]-[9].

Data Model. In our quantification and evaluation, we employ the same graph model as in [1]-[9] to represent social graphs. Specifically, the anonymized social network is modeled by graph $G^a = (V^a, E^a)$, where $V^a = \{i | i \text{ is an anonymized user}\}$ and $E^a = \{e^a_{i,j} | i, j \in V^a$, a social tie exists between i and $j\}$. To de-anonymize G^a, we use an auxiliary social network which has overlap users with G^a and can be obtained through multiple manners, e.g., data aggregation, data mining, collaborative information systems, knowledge/data brokers, etc. [1]-[9][17][18]. The auxiliary social network is also modeled by a graph $G^u = (V^u, E^u)$, where $V^u = \{i | i \text{ is a known user}\}$ and $E^u = \{e^u_{i,j} | i, j \in V^u$, an social tie exists between i and $j\}$. To conduct the theoretical quantification without involving too much mathematical details, we assume both G^a and G^u are undirected graphs. Nevertheless, our quantification can be extended to directed graphs with some straightforward technical modification. Furthermore, since our quantification and evaluation are based on the graph model, our work can be potentially applied to other kinds of data which can be modeled by graphs.

Given $i \in V^a$, its *neighborhood* is defined as $N^a_i = \{j | j \in V^a \wedge \exists e^a_{i,j} \in E^a\}$. Then, we define $d^a_i = |N^a_i|$ as the *degree* of i. Similarly, for $j \in V^u$, we can define its *neighborhood* N^u_j and *degree* d^u_j.

Graph Sampling. To make the quantification mathematically tractable, we employ the same assumptions on G^a and G^u in [6]-[9]. First, $V^a = V^u = \{1, 2, \cdots, n\}$ [6]-[9]. If $V^a \neq V^u$, we can simply satisfy this assumption by adding the users in $V^u \setminus V^a$ to V^a and adding the users in $V^a \setminus V^u$ to V^u without changing E^a or E^u. Note that this is only a mathematical assumption without limiting the generality of this work. Our quantification is also valid in the case $V^a \neq V^u$.

Second, based on the first assumption, we assume that G^a *and* G^u *are two sampling versions of an underlying conceptual graph* $G = (V, E)$ *in the physical world, where* $V = V^a = V^u$ *and* E *is the set of the true relationships among users in* V [6][8][9]. Particularly, we assume G^a is sampled from G by *independently and identically* sampling each edge in E with probability s_a, i.e., for $\forall e_{i,j} \in E$, $\Pr(e_{i,j} \in E^a | e_{i,j} \in E)$. Similarly, G^u is another sampled version of G with probability s_u. This assumption is also reasonable since people usually involve in multiple social networks and G^a and G^u are some particular social networks of users in V. For instance, G^a could be LinkedIn (a professional social network of V) while G^u is Facebook (a friendship social network of V).

De-anonymization. Based on our data model, a de-anonymization scheme can be formally defined as a mapping: $\sigma : G^a \to G^u$. Under σ, $\forall i \in V^a$, its mapping is $\sigma(i) \in V^u$. Since $V^a = V^u$, for simplicity, we define a *successful de-anonymization* of $i \in V^a$ is achieved under σ if $i = \sigma(i)$. In addition, we use σ_0 to denote the *perfect de-anonymization*, i.e., $\sigma_0 = \{(i, i) | i = 1, 2, \cdots, n\}$, and σ_k to denote any de-anonymization scheme with k incorrect mappings, i.e., k users are incorrectly de-anonymized under σ_k. Evidently, $k \in [2, n]$.

Most existing de-anonymization algorithms (e.g., [1][2][3]) consist of two phases: *seed identification phase* which identifies some *seed mapping information* from V^a to V^u and *mapping propagation phase* which propagates the seed mapping information to de-anonymize the rest of the anonymized users. In this paper, we focus on quantifying the de-anonymizability of social networks with seed knowledge. Therefore, as in [1][2][3], we assume we have identified a *seed mapping set* from V^a to V^u by some technique (e.g., the methods in [1][2][3]), denoted by $\mathcal{S} = \{(i, \sigma(i)) | i \in V^a, \sigma(i) \in V^u, i = \sigma(i)\}$. Furthermore, we define $\Lambda = |\mathcal{S}|$ as the number of seed mappings. For convenience, we denote the seed users in V^a and V^u as $\mathcal{S}^a = \{i | (i, \sigma(i)) \in \mathcal{S}\}$ and $\mathcal{S}^u = \{i | (\sigma^{-1}(i), i) \in \mathcal{S}\}$, respectively. Then, our problem now is to quantify the de-anonymizability of a social network G^a given $\mathcal{S}$, G^u, and the existing of G, s_a, and s_u.

To make the quantification easy to follow, we further assume $s_a = s_u = s$. Note that, this assumption does not change our analysis in any material detail. All our quantification results can be extended to the case $s_a \neq s_u$ only with more complex expressions.

Measuring σ**.** Given G^a, G^u, and a de-anonymization scheme σ, we measure σ by the *edge difference* between G^a and G^u under σ. First, $\forall e^a_{i,j} \in E^a$, we define $\sigma(e^a_{i,j}) = e^u_{\sigma(i),\sigma(j)}$. Furthermore, let $E^a_i(A \subseteq V^a) = \{e^a_{i,v} | v \in N^a_i \cap A\}$, and $\sigma(E^a_i(A)) = \{\sigma(e^a_{i,v}) | e^a_{i,v} \in E^a_i(A)\}$ ($\sigma(e^u_{i,j})$, $E^u_i(A)$, and $\sigma^{-1}(E^u_i(A))$ are defined in the same way). Specifically, let $E^a_i = E^a_i(V^a)$ and $E^u_j = E^u_j(V^u)$ for convenience. Then, we can define the edge difference induced by mapping $(i, \sigma(i) = j) \in \sigma$ as $\Delta_{\sigma:(i,j)} = |\sigma(E^a_i) \setminus E^u_j| + |\sigma^{-1}(E^u_j) \setminus E^a_i|$, i.e., $\Delta_{\sigma:(i,j)}$ measures the edge difference of users i and j under σ. Based on $\Delta_{\sigma:(i,j)}$, we measure σ by $\Delta_\sigma = \sum\limits_{(i,j)\in\sigma} \Delta_{\sigma:(i,j)}$, which indicates the edge difference between G^a and G^u under σ. Intuitively, since G^a and G^u are strongly correlated (highly similar), it is expected that $\Delta_{\sigma_0} \leq \Delta_{\sigma_k}$ for $k \in [2, n]$ (we demonstrate this conclusion in Sections IV and V).

Similar as $\Delta_{\sigma:(i,j)}$ and Δ_σ, we define $\Delta_{\sigma:(i,j)}(\mathcal{S})$ which measures the the edge difference of a mapping (i, j) with respect to $\mathcal{S}$: $\Delta_{\sigma:(i,j)}(\mathcal{S}) = |\sigma(E^a_i(\mathcal{S}^a)) \setminus E^u_j(\mathcal{S}^u)| + |\sigma^{-1}(E^u_j(\mathcal{S}^u)) \setminus E^a_i(\mathcal{S}^a)|$, and $\Delta_\sigma(\mathcal{S})$ which measures the edge difference of a de-anonymization scheme σ with respect to $\mathcal{S}$: $\Delta_\sigma(\mathcal{S}) = \sum\limits_{(i,j)\in\sigma} \Delta_{\sigma:(i,j)}(\mathcal{S})$.

IV. Quantification under the Erdös-Rényi Model

In this section, we quantify the de-anonymizability of G^a under the ER model, i.e., we assume $G(V, E)$ is a *random graph* $G(n, p)$, where n is the number of nodes and p specifies the probability of an edge existing between two nodes. Although real world social networks rarely satisfy the ER model [19], the analysis in this section can shed the light on the quantification in general scenarios (Section V).

A. $\mathcal{S}$ based Quantification

As a warm up, we first quantify the de-anonymizability of G^a only based on the seed information $\mathcal{S}$. For the de-anonymization scheme σ, we assume σ *de-anonymizes each user* $i \in V^a \setminus \mathcal{S}^a$ *to some user* $\sigma(i) \in V^u \setminus \mathcal{S}^u$ such that $(i, \sigma(i))$ induces the least $\Delta_{\sigma:(i,\sigma(i))}(\mathcal{S})$ [1].

We introduce a useful lemma as follows.

Lemma 1. *[6] Let X and Y be two binomial random variables with means λ_X and λ_Y, respectively. Then, when $\lambda_X > \lambda_Y$, $\Pr(X - Y \leq 0) \leq 2\exp(-\frac{(\lambda_X - \lambda_Y)^2}{8(\lambda_X + \lambda_Y)})$.*

Now, we are ready to quantify the de-anonymizability of G^a as shown in Theorem 1. We omit the proof of Theorem 1 due to the space limitation.

Theorem 1. *If $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n+1}{\Lambda}$ (i.e., $\Lambda \geq \frac{4(2\ln n+1)(2-s-ps)}{ps^3(1-p)^2}$), then it is* asymptotically almost surely (a.a.s.)[2] *that $\forall i \in V^a \setminus \mathcal{S}^a$, i is perfectly de-anonymizable under any given de-anonymization scheme σ.*

In Theorem 1, we quantify the condition on p, s, and $\mathcal{S}$ for perfectly de-anonymizing any user in $V^a \setminus \mathcal{S}^a$. Now, we quantify the condition requirement for a stronger conclusion in Theorem 2, which indicates the condition on p, s, and $\mathcal{S}$ such that all the users in $V^a \setminus \mathcal{S}^a$ are perfectly de-anonymizable. We omit the proof of Theorem 2 due to the space limitation.

Theorem 2. *If $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n+\ln(2(n-\Lambda))}{\Lambda}$ (i.e., $\Lambda \geq \frac{4(2\ln n+\ln(2(n-\Lambda)))(2-s-ps)}{ps^3(1-p)^2}$), it is a.a.s. that all the users in $V^a \setminus \mathcal{S}^a$ are perfectly de-anonymizable.*

B. Sophisticated Quantification: Considering more Structural Information

In the previous subsection, we quantified the de-anonymizablity of G^a based only on the seed knowledge. Actually, besides the edges in $E_i^a(\mathcal{S})/E_i^u(\mathcal{S})$, all the edges in E_i^a/E_i^u can provide structural information which can be used for de-anonymization. In this subsection, we consider to quantify the de-anonymizability of G^a based on all the adjacent edges of $i \in V^a$, i.e., we consider both the structural information carried by seed mappings in $\mathcal{S}$ and the overall topological information of G^a and G^u. First, we quantify the structural conditions on G^a and G^u for perfect de-anonymization in Theorem 3. Theorem 3 has two parts. The first part shows the condition such that $\Delta_{\sigma_0} < \Delta_{\sigma_k}$ for any given σ_k. The second part demonstrates the condition for a much stronger conclusion such that σ_0 is the one and the only one inducing the least edge difference. Basically, the first part of Theorem 3 can be proven using a similar technique as in [6]. Here, we obtain a tighter bound by applying more elegant quantification techniques. We omit the proof of Theorem 3 due to the space limitation.

Theorem 3. *(i) If $\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n+1}{k(n-k/2-1)}$, it is a.a.s. that $\Delta_{\sigma_0} < \Delta_{\sigma_k}$ $(k \in [2, n])$, i.e., it is a.a.s. that the perfect de-anonymization scheme σ_0 induces less edge difference than any given de-anonymization scheme $\sigma_k \neq \sigma_0$; (ii) If $\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{(k+2)\ln n+\ln(2(n-\Lambda-1))}{k(n-k/2-1)}$, it is a.a.s. that the perfect de-anonymization scheme σ_0 induces the least edge difference than all the other de-anonymization schemes, i.e., it is a.a.s. that σ_0 is the only scheme inducing the least edge difference.*

Theorem 3 has a strong implication: *even without any seed information, it still possible to perfectly de-anonymize a large scale social network*. We summarize this implication in Corollary 1.

Corollary 1. *If $\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{(k+2)\ln n+\ln(2(n-1))}{k(n-k/2-1)}$, it is a.a.s. that the perfect de-anonymization scheme σ_0 induces the least edge difference than all the other de-anonymization schemes, i.e., it is a.a.s. that σ_0 is the only scheme inducing the least edge difference.*

Based on Theorems 2, 3 and Corollary 1, it is straightforward to obtain a more accurate (tighter) bound on the structure condition of G^a and G^u for perfect de-anonymization as shown in Theorem 4.

Theorem 4. *If $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \min\{\frac{2\ln n+\ln(2(n-\Lambda))}{\Lambda}, \frac{(k+2)\ln n+\ln(2(n-\Lambda-1))}{k(n-k/2-1)}\}$, where $\Lambda \in [0, n]$, G^a is perfectly de-anonymizable.*

C. Quantification with Error Toleration

Now, we study the structural condition on G^a and G^u given $\mathcal{S}$ such that some *de-anonymization error* is tolerated. Let $\epsilon \in [0, 1 - \frac{\Lambda}{n}]$ be some constant value. We define G^a is $(1-\epsilon)$-*de-anonymizable* if at least $(1-\epsilon)n$ users in G^a are perfectly de-anonymizable. Then, we specify the condition such that G^a is $(1-\epsilon)$-deanonymizable with or without seed information in Theorem 5, i.e., the condition that at most ϵn incorrect de-anonymizations are allowable. We defer the proof to Appendix for readability.

Theorem 5. *If $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \min\{\frac{2\ln n+\ln(2(n-\epsilon n-\Lambda))}{\Lambda}, \frac{(k+2)\ln n+\ln(2(n-\epsilon n-\Lambda))}{k(n-k/2-1)}\}$, where $\Lambda \in [0, n]$, then G^a is $(1-\epsilon)$-de-anonymizable.*

V. Quantification in General Scenarios

Although the ER model is suitable to enable elegant theoretical analysis on the de-anonymizability of social networks, the fact is that it is extremely rare, if not impossible, to see

[1] Since our focus is on quantifying the de-anonymizability of G^a, we do not consider the actual de-anonymization algorithms. Specifically, we are aiming at providing the theoretical foundation on the workability of structure based de-anonymization attacks, e.g., [1][2][3].

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

real world social networks actually follow the ER model [19]. Nevertheless, the analysis under the ER model can shed light on the theoretical quantification of the de-anonymizability of social networks in general scenarios.

In this section, we quantify the de-anonymizability of G^a in general scenarios, i.e., unlike in Section IV, we assume $G(V, E)$ now could be some graph following an arbitrary network model. To accelerate the quantification, we make some definitions as follows. Given a graph $G(V, E)$ with $|V| = n$ and $|E| = m$, its *graph density* is defined as $\rho = \frac{2m}{n(n-1)}$. Let $U \subseteq V$. The *subgraph* of G on U is defined as $G[U] = G(U, E_U = \{e_{i,j} \in E | i, j \in U\})$. Furthermore, let $n_U = |U|$ and $m_U = |E_U|$. Then, the *subgraph density* of G on U is $\rho_U = \frac{2m_U}{n_U(n_U-1)}$. Let U and W be two disjoint subsets of V ($U \cap W = \emptyset$), $E_{U,W} = \{e_{i,j} \in E | i \in U, j \in W\}$ be the set of edges connecting U and W, and $m_{U,W} = |E_{U,W}|$. Then, the *connectivity* between U and W is defined as $\gamma_{U,W} = \frac{m_{U,W}}{n_U \cdot n_W}$. Finally, for the seed mapping set $\mathcal{S}$, we assume it is randomly identified. Then, a user in V is selected with probability $q = \frac{\Lambda}{n}$. We denote seed users as a set S, i.e., $S = \mathcal{S}^a = \mathcal{S}^u$. Furthermore, let set $A = V \setminus S$.

A. $\mathcal{S}$ based Quantification

In this subsection, we quantify the de-anonymizability of a social network given a seed mapping set $\mathcal{S}$. First, we show the condition for perfectly de-anonymizing an anonymized user in Theorem 6. We defer the proof to Appendix for readability.

Theorem 6. *If* $\frac{1}{4} \cdot \frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}} \geq \frac{2\ln n+1}{d_i}$, *where* $q = \Lambda/n$ *and* $\gamma_{S,A} = \frac{m_{S,A}}{\Lambda(n-\Lambda)}$, *it is a.a.s. that* $\forall i \in A$, *i is perfectly de-anonymizable.*

We further quantify the condition to perfectly de-anonymize all the users in A in Theorem 7.

Theorem 7. *If* $\frac{1}{4} \cdot \frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}} \geq \frac{2\ln n+\ln(2(n-\Lambda))}{d_i}$, *where* $q = \Lambda/n$ *and* $\gamma_{S,A} = \frac{m_{S,A}}{\Lambda(n-\Lambda)}$, *it is a.a.s. that G^a is perfectly de-anonymizable.*

B. Sophisticated Quantification: Considering more Structural Information

In the previous subsection, the perfect de-anonymizability of social networks is quantified in general scenarios based on $\mathcal{S}$. As we discussed in Section IV, for $i \in A$, besides the structural connection to the users in S, the structural information between i and other users in A is also helpful to improve the de-anonymization performance (as shown in Theorem 3). Similar to the quantification under the ER model, we quantify the de-anonymizability of social networks by considering the overall structural information in Theorem 8. We defer the proof to Appendix for readability.

Theorem 8. *(i) If* $\frac{1}{4} \cdot \frac{s^3(1-\max\{\gamma_{V_0,V_k}, \rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k}, \rho_{V_k}\}} \geq \frac{2\ln n+1}{m_{V_0,V_k}+m_{V_k}-k/2}$, *it is a.a.s. that* $\Delta_{\sigma_0} < \Delta_{\sigma_k}$ $(k \in [2, n])$, *i.e., it is a.a.s. that the perfect de-anonymization scheme σ_0 induces less edge difference than any given de-anonymization scheme $\sigma_k \neq \sigma_0$; (ii) If* $\frac{1}{4} \cdot \frac{s^3(1-\max\{\gamma_{V_0,V_k}, \rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k}, \rho_{V_k}\}} \geq \frac{(k+2)\ln n+\ln(2(n-\Lambda-1))}{m_{V_0,V_k}+m_{V_k}-k/2}$, *it is a.a.s. that the perfect de-anonymization scheme σ_0 is the only scheme inducing the least edge difference, i.e., G^a is perfectly de-anonymizable.*

Similar as Theorem 3, Theorem 8 also implies a large scale social network is perfectly de-anonymizable without seed information in general scenarios. We summarize the condition in Corollary 2.

Corollary 2. *If* $\frac{1}{4} \cdot \frac{s^3(1-\max\{\gamma_{V_0,V_k}, \rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k}, \rho_{V_k}\}} \geq \frac{(k+2)\ln n+\ln(2(n-1))}{m_{V_0,V_k}+m_{V_k}-k/2}$, *it is a.a.s. that the perfect de-anonymization scheme σ_0 is the only scheme inducing the least edge difference, i.e., G^a is perfectly de-anonymizable.*

Based on Theorems 7, 8 and Corollary 2, it is straightforward to have the following conclusion.

Theorem 9. *If* $\frac{1}{4} \cdot \frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}} \geq \frac{2\ln n+\ln(2(n-\Lambda))}{d_i}$ *or* $\frac{1}{4} \cdot \frac{s^3(1-\max\{\gamma_{V_0,V_k}, \rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k}, \rho_{V_k}\}} \geq \frac{(k+2)\ln n+\ln(2(n-\Lambda-1))}{m_{V_0,V_k}+m_{V_k}-k/2}$, *where* $\Lambda \in [0, n]$, *it is a.a.s. that G^a is perfectly de-anonymizable.*

C. Quantification with Error Toleration

Now, we quantify the $(1-\epsilon)$*-de-anonymizability* of social networks in general scenarios, where now ϵn ($\epsilon \in [0, 1 - \frac{\Lambda}{n}]$) users are allowed to be incorrectly de-anonymized. We demonstrate the quantification in Theorem 10. We defer the proof to Appendix for readability.

Theorem 10. *If (i)* $\frac{1}{4} \cdot \frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}} \geq \frac{2\ln n+\ln(2(n-\epsilon n-\Lambda))}{d_i}$ *or (ii)* $\frac{1}{4} \cdot \frac{s^3(1-\max\{\gamma_{V_0,V_k}, \rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k}, \rho_{V_k}\}} \geq \frac{(k+2)\ln n+\ln(2(n-\epsilon n-\Lambda))}{m_{V_0,V_k}+m_{V_k}-k/2}$, *where* $\Lambda \in [0, n]$, *G^a is $(1-\epsilon)$-de-anonymizable.*

VI. Large Scale Evaluation

In this section, we conduct a large scale evaluation on the de-anonymizability of 24 real world datasets based on our quantification. Our evaluation consists of two parts: *evaluation of perfect de-anonymizability* and *evaluation of* $(1-\epsilon)$*-de-anonymizability*.

A. Datasets and Setup

In the evaluation, we employ 24 various real world social datasets that mainly come from the *Stanford Large Network Dataset Collection* [20], *ASU Social Computing Data Repository* [21], and other sources [22][23]. The employed datasets are shown in Table I with preliminary statistics, where n is the number of users (nodes), m is the number of edges among users, ρ is the graph density, $\overline{d}$ is the average degree of the users, and $p(k)$ $(k = 1, 5)$ is the percentage of users with degree less than or equal to k. We further briefly introduce the datasets as follows.

Hyves *[21] is the most popular social network in the Netherlands.* ***Douban*** *[21] is a Chinese Web 2.0 site that provides user review and recommendation services for movies, books, and music.* ***Friendster*** *[21] is a social gaming site and was a social networking service website before being redesigned.* ***YouTube*** *[21] is a well known video sharing website on which users can upload, share, and view videos.*

TABLE I. DATASET STATISTICS.

Name	n	m	ρ	$\overline{d}$	$p(1)$	$p(5)$
Hyves	1,402,673	2,777,419	2.82E-06	3.96	56.76%	88.74%
Douban	154,908	327,162	2.73E-05	4.22	66.57%	90.81%
Friendster	5,689,498	14,067,887	8.69E-07	4.95	60.19%	91.27%
YouTube	1,138,499	2,990,443	4.61E-06	5.25	53.16%	85.53%
Flixster	2,523,386	7,918,801	2.49E-06	6.28	59.49%	87.26%
Last.fm	1,191,812	4,519,340	6.36E-06	7.58	47.27%	81.62%
FB-NO-wall	45,813	183,412	1.75E-04	8.01	24.18%	60.91%
Gowalla	196,591	950,327	4.92E-05	9.70	25.20%	64.50%
Foursquare	639,014	3,214,986	1.57E-05	10.06	51.10%	79.11%
Enron	33,696	180,811	3.19E-04	10.73	28.09%	67.86%
Skitter	1,694,616	11,094,209	7.73E-06	13.09	12.80%	55.41%
Slashdot	82,168	582,533	1.73E-04	14.18	2.19%	64.78%
Digg	771,229	5,907,413	1.99E-05	15.32	45.64%	77.31%
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90	20.99%	50.53%
HepPh	11,204	117,649	1.87E-03	21.00	9.95%	49.99%
AstroPh	17,903	197,031	1.23E-03	22.01	5.34%	33.69%
FB-NO-links	63,731	817,090	4.02E-04	25.64	12.71%	36.11%
Pokec	1,632,803	22,301,964	1.67E-05	27.32	10.04%	30.66%
BlogCatalog	97,884	1,668,647	3.48E-04	34.10	28.24%	59.59%
Google+	4,692,671	90,751,480	8.24E-06	38.68	5.44%	27.33%
Livemocha	104,103	2,193,083	4.05E-04	42.13	6.56%	27.56%
Twitter	456,293	12,508,272	1.20E-04	54.83	5.30%	19.76%
Orkut	3,072,441	117,185,083	2.48E-05	76.28	2.21%	7.28%
Flickr	80,513	5,899,882	1.82E-03	146.56	0.00%	11.63%

Flixster *[21] is a social movie site allowing users to share movie ratings, discover new movies and meet others with similar tastes in movies.* ***Last.fm*** *[21] is a music discovery service that gives a user personalized recommendations based on the music that user listens to.* ***Facebook-New Orleans-links (FB-NO-links) and Facebook-New Orleans-wall (FB-NO-wall)*** *[22]: Facebook is one of the most popular social networks, which connects people with friends and others who work, study, and live around them. The employed FB-NO-links dataset is a Facebook friendship network at the New Orleans area and the FB-NO-wall is a Facebook interaction (wall posts) network at the New Orleans area.* ***Gowalla*** *[20] is a location-based social networking website where users share their locations by checking-in.* ***Foursquare*** *[21] helps people to find the places to go with friends and discover food, nightlife, and entertainment for users.* ***Enron*** *[20] is an email communication dataset released by Federal Energy Regulatory Commission during its investigation.* ***Skitter*** *[20] is an Internet topology graph of* Autonomous Systems. ***Slashdot*** *[20] is a technology-related news website known for its specific user community.* ***Digg*** *[21] is a news aggregator with an editorially driven front page, aiming to select stories specifically for the Internet audience.* ***LiveJournal*** *[20] is social network for journals and blogs.* ***HepPh*** *[20] is a citation graph of the papers posted on arXiv in the high-energy physics area.* ***AstroPh*** *[20] is a collaboration network of the authors of papers posted on arXiv in the astro physics area.* ***Pokec*** *[20] is the most popular on-line social network in Slovakia.* ***BlogCatalog*** *is a social blog directory which manages the bloggers and their blogs.* ***Google+*** *[23] is one of the most popular social networking and identity services.* ***Livemocha*** *[21] is the world's largest online language learning community, offering free and paid online language courses in 35 languages.* ***Twitter*** *[21] is an online social networking and microblogging service.* ***Orkut*** *[21] is an on-line social network where users form friendship with each other.* ***Flickr*** *[21] is an image hosting and video hosting website.*

For each employed dataset, we use the raw data except for removing isolated users (most datasets do not contain any isolated users). Note that, our quantification is not limited to connected graphs. It is also applicable to disconnected social networks. Furthermore, we do not consider the direction information even if a dataset is a directed network. Again, this assumption does not limit the evaluation or quantification. Since the *direction information* can be used to improve the effectiveness of de-anonymization attacks [2], it is possible that our quantification and evaluation can be improved if we have more knowledge, e.g., the direction information. One of the future works is to quantify the de-anonymizability of directed social networks.

In our quantification, the seed mappings are chosen randomly, i.e., the high-degree users are not given preference as in [8][9] although they may be more helpful as seed mappings. Consequently, our evaluation results represent the general results of our quantification. Nevertheless, our evaluation demonstrate that most, if not all, of the social datasets are partially or perfectly de-anonymizable based on their structural information. We quantitatively show how many users can be successfully de-anonymized in each social network (i.e., $1-\epsilon$) in our evaluation.

We quantify the de-anonymizability of a social network using seed information and using the overall structural information, respectively. Therefore, we use suffixes "-S" and "-A" to distinguish these two scenarios (e.g., Twiiter-A and Twitter-S), where "-S" and "-A" imply using seed information and overall structural information, respectively. *If we do not specify the suffix or the particular context, it implies using the overall structural information by default.* Furthermore, due to the space limitation, we do not show all the experimental results.

B. Evaluation of Perfect De-anonymizability

In this section, we evaluate the condition for perfect de-anonymizability of the datasets in Table I.

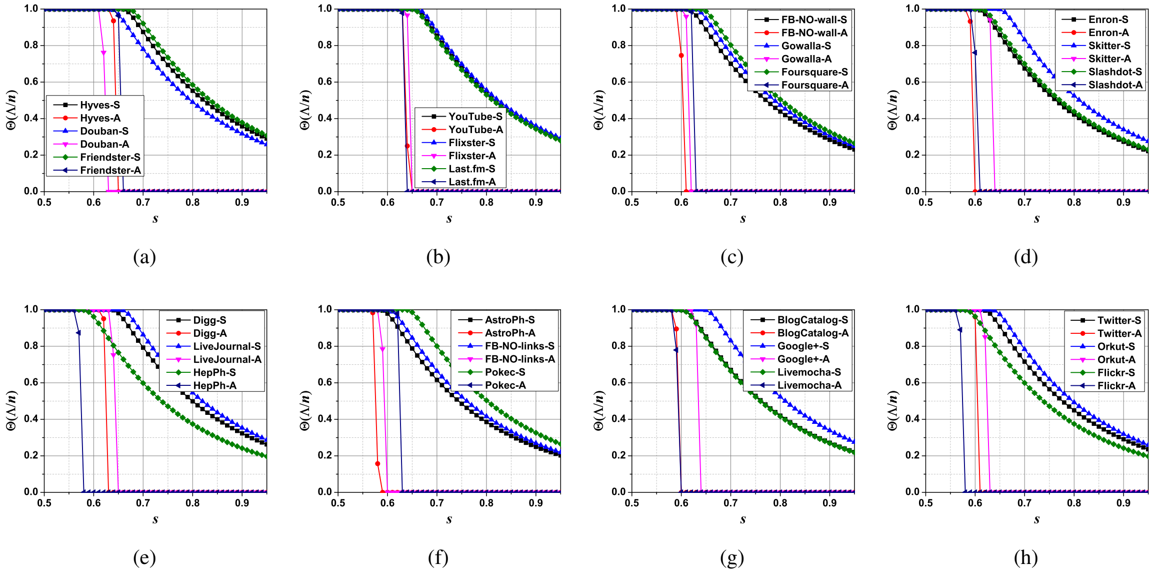
1) Evaluation on Λ: Based on our quantification, we evaluate the requirements on the size of seed mappings Λ and the sampling rate s for the perfect de-anonymizability of each dataset in Fig.1. Since all the datasets have different sizes, for convenience, we show $\Theta(\Lambda/n)$ instead of Λ directly.

From Fig.1, we have the following observations.

(i) If the overall structural information is considered, each dataset is *asymptotically perfectly de-anonymizable*[3] even without any seed information when s is above some threshold value, which is consistent with our theoretical quantification[4]. For instance, the Twitter dataset is asymptotically perfectly de-anonymizable when $s \geq 0.61$ without seed information if the overall structural information is considered. This implies that the structure itself is sufficient to break the privacy. The reason for this result is that the perfect de-anonymization scheme induces the least edge difference as shown in our quantification.

[3] To be accurately, *asymptotically perfectly de-anonymizable* here implies $\Theta(n)$ users of each dataset can be successfully de-anonymized.

[4] Actually, the quantification does not implies a computationally efficient algorithm. It is still an open problem to find an efficient (polynomial-time) algorithm with provable performance guarantee.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 1. Perfect de-anonymization: $\Theta(\Lambda/n)$ vs. s. Since the quantification (Theorem 9) for perfect de-anonymization is meaningful for large n, we set $n = 1000/\rho$ for each social network in this group of evaluations. All the other network properties, e.g., ρ, $\overline{d}$, degree distribution, etc., remain the same as in the original dataset.

(ii) If s is below some threshold value, it is necessary to have $\Theta(n)$ seed mappings to perfectly de-anonymize each social network, i.e., each social network is not perfectly de-anonymizable unless $\Theta(n)$ users are identified as seeds. For instance, Google+ is not perfectly de-anonymizable when $s < 0.61$. The reason is that a small s implies less edges are sampled into G^a and G^u. It follows that most of the users are low degree users and thus the structural information is not sufficient to achieve perfect de-anonymization.

(iii) For the de-anonymization only based on seed information ("*-S"), to achieve perfect de-anonymization, the required number of seed mappings decreases when s increases as expected. For instance, if only the seed information is considered to perfectly de-anonymize Google+, 49.27% seed users are needed when $s = 0.8$ while 31.89% seed users are needed when $s = 0.9$. This is because a large s implies more structural similarity between G^a and G^u. Thus, less seed mappings are needed to distinguish all the users.

(iv) Given some s, a social network with higher graph density requires fewer seed mappings. For example, to be perfectly de-anonymizable, 49.27% seed users are required for Orkut-S ($\rho = 2.48$E-5) while 37.47% seed users are required for Flickr-S ($\rho = 1.82$E-3). This is also true for the overall structural information based de-anonymization. This is because a higher graph density implies that more structural information is carried by the data, followed by more structural information can be used to distinguish users.

Now, we examine the behavior of Λ when we fix the graph density of each social network while varying n. The results are shown in Fig.2, where $x \times n$ (the x-axis) denotes the number of users is x times of the original size n.

From Fig.2, we have the following observations.

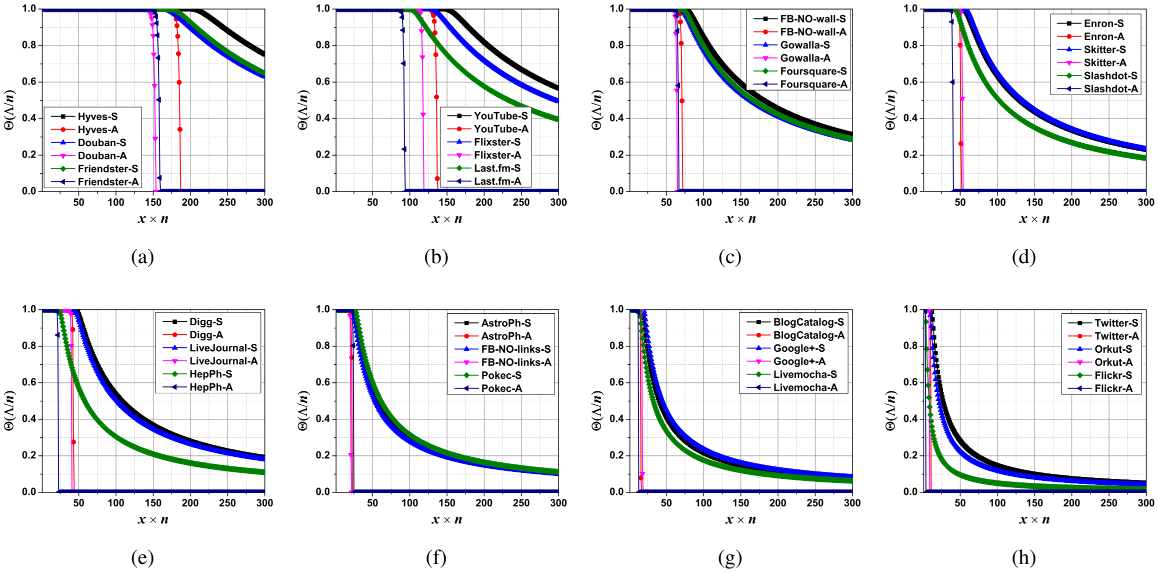
(i) When n is above some threshold value, each social network is perfectly de-anonymizable based on the overall structural information, which confirms the conclusion of Theorem 8. The reason is straightforward. More structural information will be available when n increases and ρ is fixed. Consequently, more users can be de-anonymized based on the structural information. Because of the same reason, for the seed based de-anonymization, the required number of seed mappings decreases when n increases.

(ii) Given $\Theta(\Lambda/n)$, the required threshold value on n is smaller for social networks with high graph densities and vice versa. This is because a high ρ implies more structural information is available followed by more similarity between G^a and G^u when s is fixed.

(iii) Similar to the scenario of changing s, when n is below some threshold value, it is necessary to have $\Theta(n)$ seed user mappings to perfectly de-anonymize a social network. This can be seen from our quantification: it is a.a.s. that σ_0 induces the least edge difference when the required condition holds and $n \to \infty$, i.e., n should be large enough.

2) Evaluation on n: In this subsection, we study the condition on n to perfectly de-anonymize a social network given different s or Λ. The objective of this group of evaluation is to study the asymptotic behavior of n in different scenarios, since our quantification is mathematically meaningful when n is a large number. Furthermore, based on our quantification, when $n \to \infty$, the overall structure based quantification will dominate the perfect de-anonymizability of a social network (this claim can be confirmed by the evaluation results in Fig.4). Consequently, we consider the overall structural information (including seed mappings) in the evaluation of n. When s is changed from 0.5 to 0.95, the requirement on the lower bound of n, i.e., $\Omega(n)$, is shown in Fig.3. From Fig.3, we have two main observations as follows.

(i) When s increases, $\Omega(n)$ decreases, e.g., to perfectly de-anonymize Google+, $\Omega(n)$ decreases from 2.85E8 to 3.19E7



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2. Perfect de-anonymization: $\Theta(\Lambda/n)$ vs. n. Default setting: $s = 0.7$.

when s increases from 0.5 to 0.9. This is because a large s implies more similarity between G^a and G^u since they share more common edges. Consequently, the condition on $\Omega(n)$ to perfectly de-anonymize a social network becomes loose. This observation can also be explained by our quantification. From Theorem 9, a large s implies a large value on the left hand of each condition, followed by smaller n requirement.

(ii) The graph density has positive influence on $\Omega(n)$, i.e., a social network with high graph density requires loose condition on $\Omega(n)$ for perfect de-anonymization, which is consistent with our quantification. For instance, Orkut ($\rho = 2.48$E-5) requires a smaller $\Omega(n)$ than Google+ ($\rho = 8.24$E-6). The reason is that a large ρ implies more structural information is carried by the dataset. Therefore, it is much easier to perfectly de-anonymize this dataset.

Now, we want to study the impact of Λ on $\Omega(n)$ to perfectly de-anonymize a social network. The results are shown in Fig.4. From Fig.4, we have the following two observations.

(i) When Ω (i.e., $\Theta(\Omega/n)$) increases, $\Omega(n)$ only has a very slight decrease, e.g., the $\Omega(n)$ of Friendster, Skitter, LiveJournal, etc., which is confirmed by our quantification that considering the overall structural information will dominate the perfect de-anonymizability of a social network when n is large ($n \to \infty$, see Theorem 9). This is because given that $\Omega(n)$, the overall structural information based de-anonymization has already been achieved. However, even the seed information based de-anonymization can de-anonymize a large portion of each social network given the same $\Omega(n)$ (as shown in Fig.1 and Fig.2), more seed mappings are necessary to perfectly de-anonymize all the users.

(ii) Again, the graph density has positive influence on $\Omega(n)$ in different settings of Λ. The reason is the same as explained before: a large ρ implies more similarity between G^a and G^u when s is fixed.

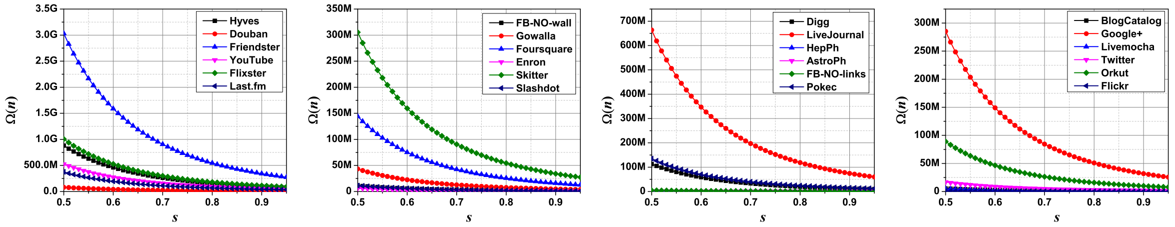
C. Evaluation of $(1-\epsilon)$-De-anonymizability

1) Evaluation on $(1-\epsilon)$: In this subsection, we evaluate the actual de-anonymizability of the 24 real world datasets by quantitatively demonstrating $(1-\epsilon)$ (note that, ϵn is the error tolerated during the de-anonymization process), i.e., *how many users in each social network can be successfully de-anonymized in each specific scenario*.

When all the structural information (including seed mappings) are considered, *the lower bound on the percentage of de-anonymizable users in the 24 social networks*, i.e., $\Omega(1-\epsilon)$, is shown in Fig.5 with different s. From Fig.5, we have the following important observations.

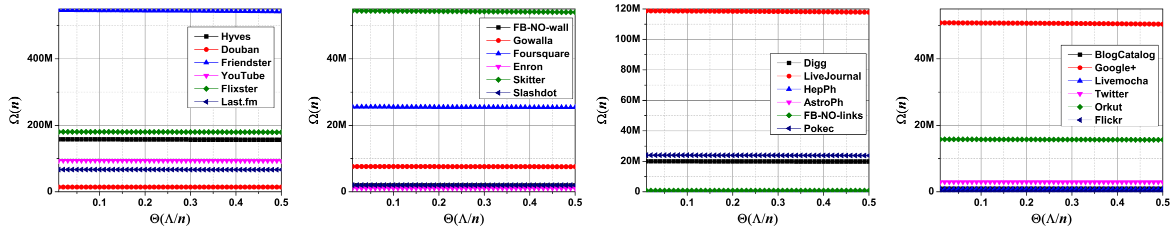
(i) All the 24 social networks are partially de-anonymizable although they may not be perfectly de-anonymizable. For instance, when $s = 0.55$, 20.88% YouTube users, 33.62% Foursquare users, 66.69% Facebook users at New Orleans, 72.94% Google+ users, and 97.6% Twitter users are perfectly de-anonymizable based on the overall structural information. Consequently, the obtained quantitative results confirmed the success of existing heuristic algorithms [2][3]. This is also consistent with our quantification on $(1-\epsilon)$-de-anonymization: *if the low-degree users are treated as the tolerated de-anonymization errors, the high-degree users are more likely to be successfully de-anonymized*, i.e., these social networks are partially de-anonymizable. In other words, when perfect de-anonymization is not achievable, these high-degree users are still perfectly de-anonymizable since they carry sufficient structural information.

(ii) When s increases, $\Omega(1-\epsilon)$ also increases, i.e., more and more users can be successfully de-anonymized for each social network. For instance, when s changes from 0.5 to 0.65, the percentage of de-anonymizable users of Google+ increases from 58.76% to 99%. The reason is similar as explained in the previous subsection: a large s implies more common edges shared by G^a and G^u, i.e., more structural similarity between G^a and G^u. Consequently, it is more likely that the



(a) (b) (c) (d)

Fig. 3. Perfect de-anonymization: n vs. s. Default setting: $\Lambda/n = 0.015$ (1.5% users are randomly chosen as seed mappings).



(a) (b) (c) (d)

Fig. 4. Perfect de-anonymization: n vs. Λ. Default setting: $s = 0.8$.

correct user de-anonymization induces less edge difference (de-anonymization error). This is also consistent with our theoretical quantification.

(iii) When s is increased above some value, several social networks can be asymptotically perfectly de-anonymized ($\Theta(n)$ users can be successfully de-anonymized). For instance, when $s \geq 0.78$, $s \geq 0.66$, and $s \geq 0.63$, over 99% users of Slashdot, FB-NO-link, and Google+ can be successfully de-anonymized, respectively. This fact comes from the same reason as the previous observation: a large s implies more structural similarity followed by more de-anonymizable a social network.

(iv) The social networks with higher average degree $\overline{d}$ is more de-anonymizable, e.g., when $s = 0.6$, 53.23% LiveJournal users ($\overline{d} = 17.9$) are de-anonymizable while 73.38% Pokec users ($\overline{d} = 27.32$) are de-anonymizable. The reason is evident: a higher $\overline{d}$ implies more common edges in G^a and G^u. Therefore, the correct de-anonymization is more likely inducing less edge difference.

Now, we study the $(1-\epsilon)$-de-anonymizability of the 24 social networks when we fix the network density, s, and Λ/n while changing n. The results are shown in Fig.6. From Fig.6, we have the following observations.

(i) When n increases, the percentage of de-anonymizable users also increases for both seed based and overall structure based de-anonymization. For instance, when the network size changes from $10n$ to $20n$, the percentage of de-anonymizable Flickr users increases from 41.65% to 59.08% in seed based de-anonymization; similarly, when network size is $5n$, 67.81% of LiveJournal users are de-anonymizable while when the network size is above $10.5n$, LiveJournal is asymptotically perfectly de-anonymizable. This fact is consistent with our quantification. The reasons is that a large n implies richer structural information when ρ is fixed. Hence, more users are de-anonymizable.

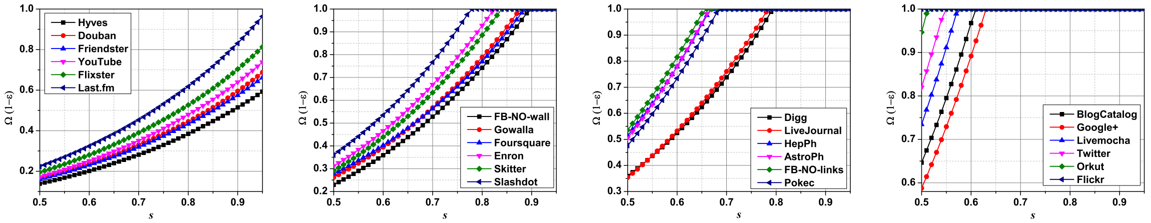
(ii) As expected, the overall structural information is more powerful in de-anonymizing social networks. This is also consistent with our quantification. Since more structural information is considered, the probability that the correct de-anonymization induces more edge differences than incorrect de-anonymization will be decreased. Consequently, "*-A" de-anonymizes more users than "*-S".

(iii) As validated before, graph density also has positive impact on $\Omega(1-\epsilon)$, i.e., a social network with a higher graph density is more de-anonymizable. The reason still comes from the fact that a high ρ implies more structural similarity between G^a and G^u.

Intuitively, if we have more seed mappings, more users should be de-anonymizable even we do not consider the overall structural information. Theoretically, this intuition is quantified in Theorem 7. We evaluate this quantification by studying the impacts of the number of seed mappings on the percentage of de-anonymizable users. The results are shown in Fig.7. From Fig.7, we have the following observations.

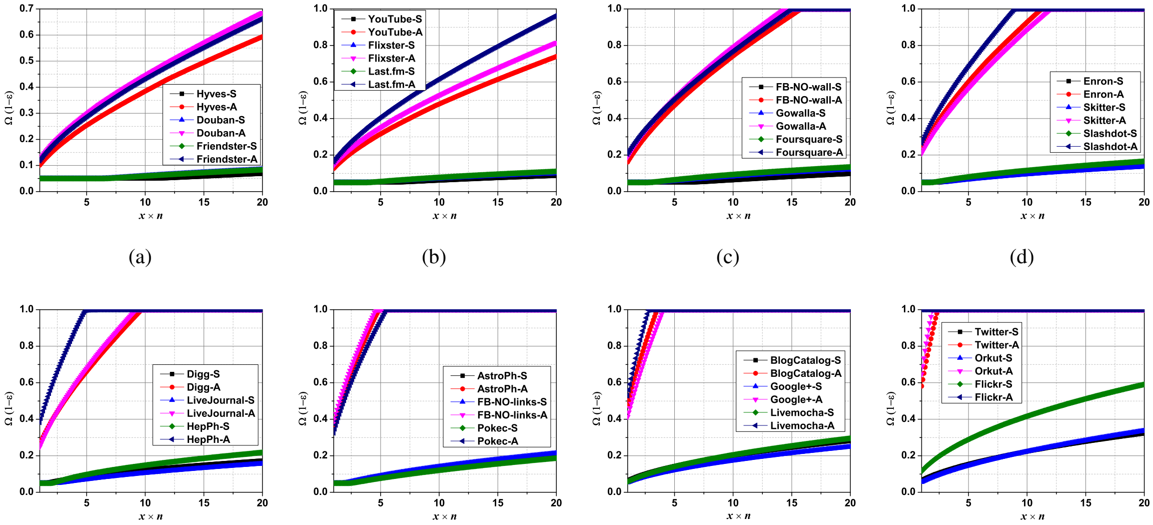
(i) When more seed mappings are available, more users are de-anonymizable, e.g., when $\Omega(\Lambda/n)$ changes from 0.05 to 0.15, the percentage of de-anonymizable Google+ users increases from 40.07% to 72.28%. The reason is evident since more seed mappings implies more knowledge is available to improve the de-anonymization accuracy, which can also be seen from our quantification.

(ii) Although ρ and $\overline{d}$ have a positive influence on $\Omega(1-\epsilon)$, it is still possible that a social network with smaller ρ or $\overline{d}$ be more de-anonymizable than a social network with higher ρ or $\overline{d}$ in some cases, e.g., BlogCatalog has a smaller $\overline{d}$ while larger ρ than Google+, and Orkut has a smaller ρ while larger $\overline{d}$ than BlogCatalog. This is because the seed mappings



(a) (b) (c) (d)

Fig. 5. $(1-\epsilon)$-de-anonymization: $\Omega(1-\epsilon)$ vs. s. Default setting: $\Lambda = 0.05n$ (5% users are seeds).



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 6. $(1-\epsilon)$-de-anonymization: $\Omega(1-\epsilon)$ vs. n. Default setting: $s = 0.8$ and $\Lambda/n = 0.05$.

in seed based de-anonymization are randomly identified and the de-anonymization process is also affected by the degree distribution of the social network. Consequently, both ρ and $\overline{d}$ have impacts on the de-anonymizability of a social network. However, it is difficult to determine which one will dominate the de-anonymizability. Generally speaking, the richer the structural information, i.e., the higher ρ and $\overline{d}$, the more de-anonymizable the social network.

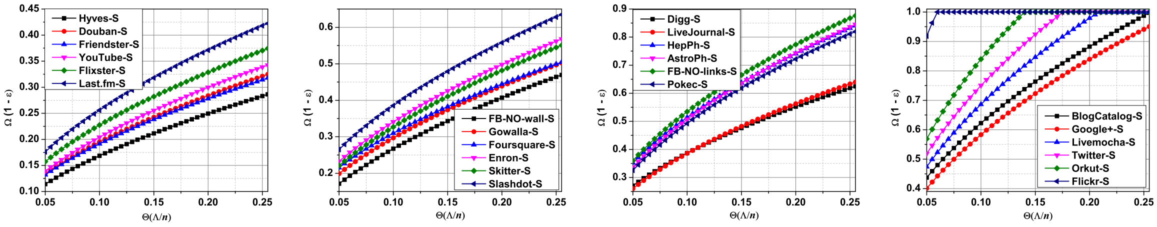
2) Evaluation on Λ: In this subsection, we evaluate the condition on Λ in $(1-\epsilon)$-de-anonymization. When $\epsilon = 0.4$, i.e. up to 40% user de-anonymization error is tolerable, the condition on Λ to perfectly de-anonymize at least $1-\epsilon = 60\%$ users of each social network under different settings of s is shown in Fig.8. From Fig.8, we can observe that:

(i) When s is below some threshold value, $\Theta((1-\epsilon)n)$ seed mappings are necessary to perfectly de-anonymize $(1-\epsilon)n$ anonymized users. For instance, when $s < 0.72$, $\Theta(\Lambda/n) \sim 0.6$ for Google+ in seed based de-anonymization, i.e., almost 60% Google+ users have to be identified as seeds; similarly, when $s < 0.51$, the condition on Λ is also $\Theta(\Lambda/n) \sim 0.6$ for Google+ in overall structural information based de-anonymization. This is because when s is small, less common edges are shared by G^a and G^u. Consequently, it tends to involve all the anonymized users as seeds to achieve perfect de-anonymizability.

(ii) For seed based de-anonymization, when s is above some threshold value, $\Theta(\Lambda/n)$ decreases with the increases of s (less seed mappings are needed), e.g., when s is increased from 0.8 to 0.9, $\Theta(\Lambda/n)$ decreases from 0.47 to 0.3. For overall structural information based de-anonymization, when s is above some value, it is a.a.s. that a social network is $(1-\epsilon)$-de-anonymizable even without any seed mapping information, e.g., when $s \geq 0.51$, $\Theta(\Lambda/n) \sim 0$ for Google+. This is because: first, when s increases, G^a and G^u are more structurally similar. Thus, G^a is more de-anonymizable in both seed and overall structural information based de-anonymization; and second, when the overall structural information is considered, the perfect de-anonymization scheme tends to induce the least edge difference when s is above some threshold value, i.e., a social network becomes $(1-\epsilon)$-de-anonymizable when s is large enough, which is also consistent with our quantification.

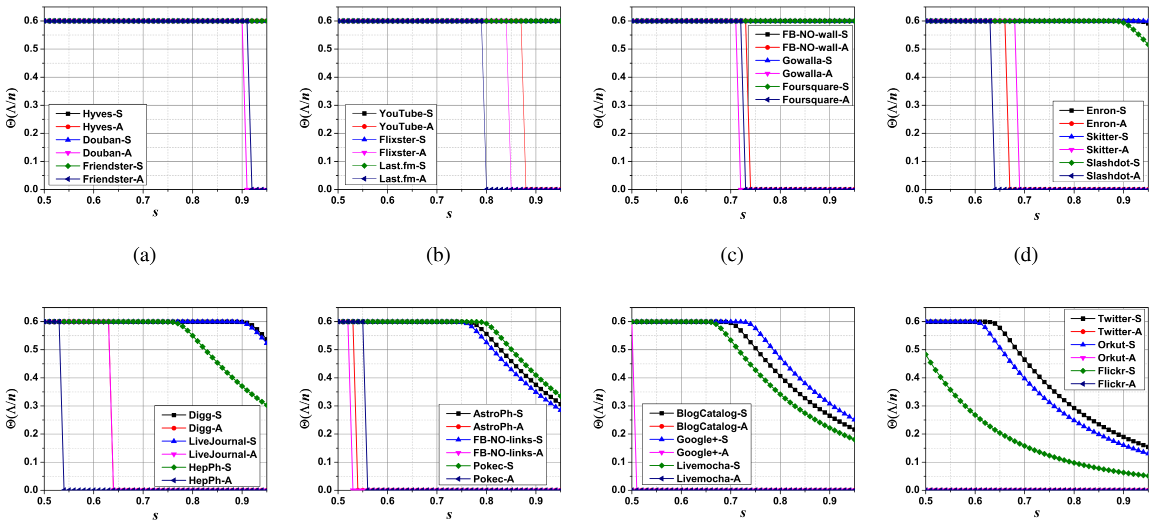
If we fix $s = 0.8$, the condition on Λ to make each social network $(1-\epsilon)$-de-anonymizable under different ϵ is shown in Fig.9. From Fig.9, we can see that:

(i) In seed based de-anonymization, to make the social networks with low average degree $(1-\epsilon)$-de-anonymizable, it is necessary to identify $\Theta((1-\epsilon)n)$ seed mappings. For example, the social networks shown in Fig.9 (a)-(d) have



(a) (b) (c) (d)

Fig. 7. $(1-\epsilon)$-de-anonymization: $\Omega(1-\epsilon)$ vs. Λ. Default setting: $s = 0.8$



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 8. $(1-\epsilon)$-de-anonymization: Λ vs. s. Default setting: $\epsilon = 0.4$.

$\overline{d} < 15$ and the condition on Λ to make them $(1-\epsilon)$-de-anonymizable is $\Theta(\Lambda/n) \sim 1-\epsilon$. The reason is that a low $\overline{d}$ implies less edges from anonymized users to seed users. Consequently, more seed mappings are necessary. On the other hand, if a social network has a large $\overline{d}$, e.g., most of the social networks in Fig.9 (e)-(h), less seed mappings are needed to be $(1-\epsilon)$-de-anonymizable in seed based de-anonymization. For instance, when $\epsilon = 0.6$, to make Google+ 0.4-de-anonymizable, 22.52% seed users are needed.

(ii) In overall structural information based de-anonymization, if ϵ (the tolerated de-anonymization error) is above some threshold value, all the 24 social networks are $(1-\epsilon)$-de-anonymizable except for Hyves, which has a very low $\overline{d} = 3.96$. The reason is that when the overall structural information (including seed mappings) is considered and $s = 0.8$, the correct de-anonymization induces the least edge difference with higher probability than in the seed based de-anonymization, which is consistent with our quantification. Again, the results also confirmed that the overall structural information based de-anonymization is more effective.

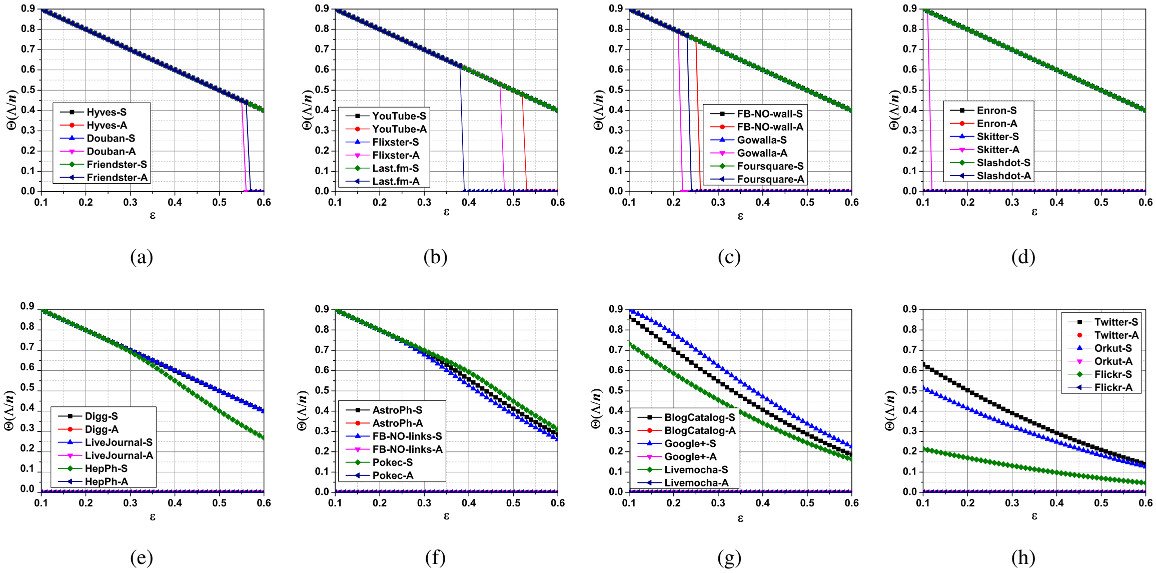
Now, we evaluate the condition on Λ when the network size changes. The results are shown in Fig.10. From Fig.10, we have the following observations.

(i) When n varies, the behavior of $\Theta(\Lambda/n)$ is similar to that when s varies. For the social networks with low $\overline{d}$, e.g., the social networks shown in Fig.10 (a)-(d), it is necessary to have $\Theta(\Lambda/n) \sim 1-\epsilon$ in seed based de-anonymization. The reason is also similar to that presented in the earlier analyis. A small $\overline{d}$ implies less edges between anonymized users and seed users. Hence, it is necessary to have $\Theta(\Lambda/n) \sim 1-\epsilon$ to perfectly de-anonymize $(1-\epsilon)n$ users. On the other hand, when the network size is above some threshold value and continue to increase, less seed mappings are needed for social networks with high $\overline{d}$ (social networks in Fig.10 (e)-(h)) to be $(1-\epsilon)$-de-anonymizable. The reason is also similar to that presented in the earlier analysis.

(ii) Again, the overall structural information based de-anonymization is more powerful, i.e., even without seed information, the structure itself can make a social network perfectly de-anonymizable. The quantification along with the evaluation results provide the foundation of the de-anonymization attacks without seed information.

VII. Discussion, Limitations, and Future Work

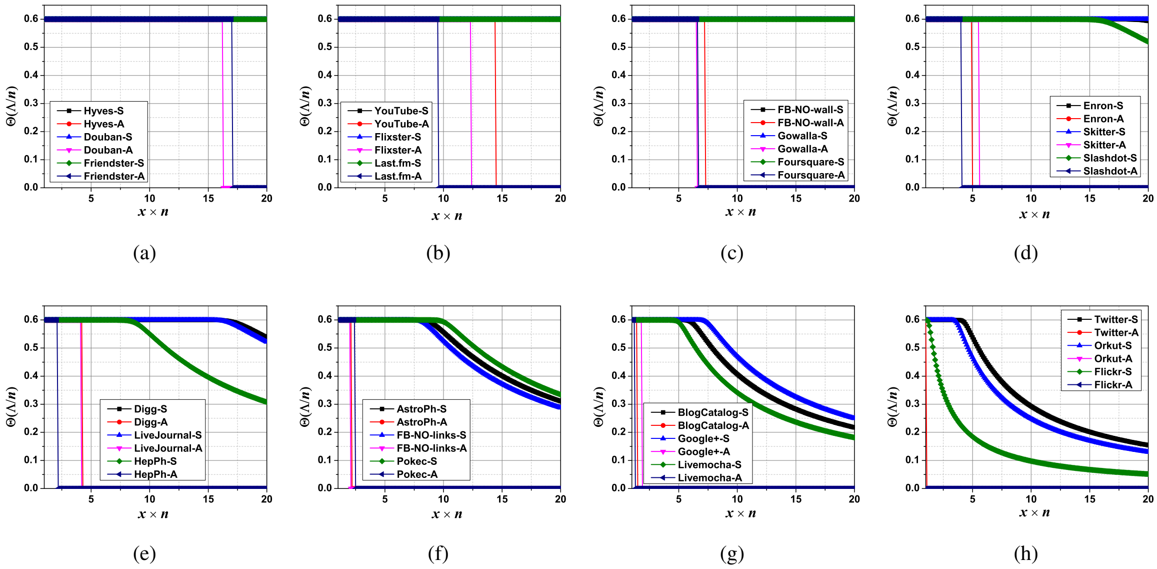
Discussion. In this paper, to the best of our knowledge, we conduct the first comprehensive theoretical quantification on the de-anonymizability of social networks under both the theoretical ER model and the general arbitrary network model.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 9. $(1-\epsilon)$-de-anonymization: Λ vs. ϵ. Default setting: $s = 0.8$.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 10. $(1-\epsilon)$-de-anonymization: Λ vs. n. Default setting: $s = 0.8$ and $\epsilon = 0.4$.

The most meaningful significance of our quantification is that it provides the theoretical foundation for the existing *de-anonymization attacks with seed information* [2][3].

An interesting implication of our quantification and evaluation is that *the overall structural information based de-anonymization is more powerful and it can perfectly de-anonymize a social network even without any seed information*. This finding lies the theoretical foundation of a further implication that *unlike the de-anonymization attacks proposed in [2][3] which depend on seed mappings, one can design new effective de-anonymization attacks without seed information*. We believe this would shed the light of future research in the social networks/structural data de-anonymization area.

Another concern is to what extent of privacy protection can existing anonymization techniques provide (e.g., *naive ID removal*, *k-anonymization* [11][12], and *differential privacy* [13][14]). As we discussed in Section II, the *naive ID removal* technique is vulnerable to the existing de-anonymization attacks [1][2][3]. For *k-anonymization*, since it is based on data's syntactic property, it cannot protect users' privacy against structure based de-anonymization even if it is satisfied [2]. Furthermore, k-anonymization is not scalable, which implies it is computationally infeasible to be extended to large scale social networks. Finally, the *differential privacy* technique is suitable for *interactive queries*, where quantified noise can be added by data owners before the data is returned/transferred to users who submitted the query (i.e., the adversary in a de-anonymization attack) [13][14]. However, the de-anonymization attacks we studied in this paper belong to the *non-interactive query* *category*. Recent efforts have proposed using *differentially private graph models* [15], however, the *differential privacy parameter* still has to be determined before data sharing, which makes this method ineffective in defending against structure based de-anonymization attacks [2][3]. In summary, it is expected that new anonymization techniques against the structural information based de-anonymization attacks will be designed. Our quantification and evaluation can shed light to the expected design by demonstrating the theoretical foundation of such attacks and their efficacy on attacking real world social networks.

Limitations. There are still some limitations of this paper.

(i) To be accurate, we consider both the edges from anonymized users to seed users and the edges among anonymized users in the quantification of the overall structural information based de-anonymization. Actually, some other global graph properties are also helpful in improving structure based de-anonymization attacks, e.g., the *betweenness/closeness centrality* of a user, the distance from a user to other users. Although we believe these graph properties can be used in improving de-anonymization attacks, it is difficult to involve them in the theoretical quantification. The reason is as follows: all these graph properties represent a user's *global topological importance/characteristics with respect to the entire graph*. Consequently, even if there is just one edge-change, it may change the global topological characteristics (e.g., betweenness/closeness centrality, the distance from an anonymized user to a seed) of an arbitrary number of users. It follows that it is very difficult, if not impossible, to quantify the change on the global topological characteristics of a user. Even though these global topological characteristics are not considered in our quantification, we demonstrate that the neighboring edges are sufficient to perfectly or partially de-anonymize a social network.

(ii) In this paper, we focus on closing the gap between existing de-anonymization practices (i.e., heuristic algorithms) and their theoretical foundation by quantifying the perfect de-anonymizability and $(1-\epsilon)$-de-anonymizability of social networks. We do not specifically consider how to design structural data anonymization techniques to defend against such de-anonymization attacks. Actually, this is still an important open problem since we have an increasing amount of social data. We believe our quantification and evaluation in this paper can shed light on the future research in this area by providing the theoretical foundation of structure based de-anonymization attacks and their effectiveness in attacking real world social networks. Furthermore, our quantification and evaluation are expected to draw the attention of data owners and help them develop more proper policies to protect social data.

Future Work. The future work of this paper will take the following directions: (i) It is expected that a new mathematical model under which we can theoretically analyze the change on users' global topological properties (e.g., betweenness/closeness centrality, the distance to seed users) will be developed. Then, we can quantify the de-anonymizability of social networks more accurately. (ii) In our quantification, we do not explicitly involve the noise level since we do not involve a specific *noise description model* (actually, to the best of our knowledge, we currently do not have proper schemes to add noise with data utility preservation). In the future, we propose to quantify the de-anonymizability of social networks by involving a function describing the existing noise. (iii) As pointed out before, we do not have effective data anonymization techniques against structure based de-anonymization attacks. We propose to study this open problem based on our quantification and evaluation and develop a *secure data publishing platform* which can examine the data de-anonymizability, anonymize data properly with utility preservation, and publish data securely. We also propose to develop new social data protection policies for data owners.

VIII. Conclusion

In this paper, we study the de-anonymizability of social networks based only on their structural information. We quantify the *perfect* and $(1-\epsilon)$-*de-anonymizability* of social networks with seed information in general scenarios, where a social network can follow an arbitrary network model. To the best of our knowledge, this is the first comprehensive quantification study on the perfect and partial de-anonymizability of social networks under a general model. Based on our quantification, we conduct a large scale evaluation on the de-anonymizability of 24 various real world social networks. Finally, we discuss the implications of this work. Our findings are expected to shed light on the future research in the structural data anonymization and de-anonymization areas, and help data owners evaluate the vulnerability of their structural data before data publishing.

Acknowledgment

The authors are grateful to Shukun Yang and Xiaojing Liao for helpful discussions. The authors are also very grateful to the anonymous reviewers for their time and valuable comments.

This work was partly supported by NSF-CAREER-CNS-0545667. Neil Zhenqiang Gong was supported by the National Science Foundation under Grant No. 0831501 CTL and by the Office of Naval Research under MURI Grant No. N000140911081. Prateek Mittal was supported in part by the NSF under the grant CNS-1409415.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

[14] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang *Membership Privacy: A Unifying Framework For Privacy Definitions*, CCS 2013.

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

[16] A. Meyerson and R. Williams, *On the Complexisty of Optimal k-Anonymity*, PODS 2004.

[17] C. Shah, R. Capra, and P. Hansen, *Collaborative Information Seeking*, Computer, Vol. 47, No. 3, pp. 22-25, 2014.

[18] 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, 2014.

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

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

[21] ASU Social Computing Data Repository, *http://socialcomputing.asu.edu/pages/datasets*.

[22] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, *On the Evolution of User Interaction in Facebook*, WOSN 2009.

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

[24] H. Yu , C. Shi , M. Kaminsky , P. B. Gibbons , and F. Xiao *DSybil: Optimal Sybil-Resistance for Recommendation Systems*, S&P 2009.

Appendix

***Proof Sketch of Theorem 5*:** First, we prove that if $\frac{1}{4}\cdot\frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n+\ln(2(n-\epsilon n-\Lambda))}{\Lambda}$, G^a is $(1-\epsilon)$-de-anonymizable. Let $V_c \subseteq V^a \setminus \mathcal{S}^a$ and $|V_c| = n - \epsilon n - \Lambda$. Furthermore, let $\mathbf{E}$ be the event that *there is at least one incorrectly de-anonymized user in V_c*. Then, using the similar proof technique in Theorem 2, we have $\Pr(\mathbf{E}) \leq \sum_{i\in V_c}\Pr(i \text{ is incorrectly de-anonymized}) \leq \sum_{i\in V_c} 2\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}\Lambda) \leq 2(n-\epsilon n-\Lambda)\exp(-2\ln n - 2\ln(2(n-\epsilon n-\Lambda))) = \frac{1}{n^2}$. Hence, it a.a.s. that $\Pr(\mathbf{E}) \sim 0$ as $n \to \infty$, i.e., it a.a.s. that G^a is $(1-\epsilon)$-de-anonymizable.

Second, we prove if $\frac{1}{4}\cdot\frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{(k+2)\ln n+\ln(2(n-\epsilon n-\Lambda))}{k(n-k/2-1)}$, G^a is also $(1-\epsilon)$-de-anonymizable. Now, we do not have to distinguish σ_0 and σ_k when $k \leq \epsilon n$. Let $\mathbf{E}$ now be the event that *there is some de-anonymization scheme σ_k such that $\sigma_k \neq \sigma_0$, $\Delta_{\sigma_k} \leq \Delta_{\sigma_0}$, and $k > \epsilon n$*. Then, we have $\Pr(\mathbf{E}) = \bigcup_{k=\epsilon n+1}^{n-\Lambda}\Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) \leq \sum_{k=\epsilon n+1}^{n-\Lambda} n^k\cdot 2\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}(m_k - k/2)) \leq \frac{1}{n^2}$. Consequently, it a.a.s. that $\Pr(\mathbf{E}) \sim 0$ as $n \to 0$, i.e., G^a is $(1-\epsilon)$-de-anonymizable. □

***Proof Sketch of Theorem 6*:** To prove this theorem, it is sufficient to prove that $\forall i \in A$, $\Delta_{\sigma:(i,i)}(\mathcal{S}) < \Delta_{\sigma:(i,j\neq i)}(\mathcal{S})$ under any given σ (it follows that i is perfectly de-anonymizable in terms of $\Delta_{\sigma:(i,\sigma(i))}(\mathcal{S})$). Let $X = \Delta_{\sigma:(i,j\neq i)}(\mathcal{S})$ and $Y = \Delta_{\sigma:(i,i)}(\mathcal{S})$ be two random variables. We have $X \underset{\text{stochastically}}{\sim} \mathbf{B}(d_i q, 2s(1-s\gamma_{S,A})), Y \sim \mathbf{B}(d_i q, 2s(1-s))$. Applying Lemma 1, we have $\Pr(X \leq Y) \leq 2\exp(-\frac{(\lambda_X-\lambda_Y)^2}{8(\lambda_X+\lambda_Y)}) = 2\exp(-\frac{1}{4}\frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}}d_i) \leq \frac{1}{n^2}$. Since $\sum_{n>0}\frac{1}{n^2} = \frac{\pi^2}{6} < \infty$, it is a.a.s. that i is perfectly de-anonymizable as $n \to \infty$. □

***Proof Sketch of Theorem 8*:** (i) Let $V_k \subseteq V\backslash S$ be the set of incorrectly de-anonymized users under $\sigma_k \neq \sigma_0$ and $V_0 = V \setminus V_k$. Furthermore, let $X = \Delta_{\sigma_k}$ and $Y = \Delta_{\sigma_0}$ be two random variables. Then, we have $Y \sim \mathbf{B}(m, 2s(1-s))$. Furthermore, we can consider four cases to quantify X. First, the edge difference caused by the edges in E_{V_0} follows $\mathbf{B}(m_{V_0}, 2s(1-s))$; second, the edge difference caused by the edges in E_{V_0,V_k} stochastically follows $\mathbf{B}(m_{V_0,V_k}, 2s(1-s\gamma_{V_0,V_k}))$; third, the edge difference caused by the non-transposition edges in E_k stochastically follows $\mathbf{B}(m_{V_k}-x, 2s(1-s\rho_{V_k}))$, where x here is the number of transposition edges under σ_k; and finally, the edge difference caused by the transposition edges in E_k follows $\mathbf{B}(x, 2s(1-s))$. Since $x \leq k/2$, we have $X \underset{\text{stochastically}}{\geq} \mathbf{B}(m_{V_0}, 2s(1-s)) + \mathbf{B}(m_{V_0,V_k}, 2s(1-s\gamma_{V_0,V_k})) + \mathbf{B}(m_{V_k} - k/2, 2s(1-s\rho_{V_k})) + \mathbf{B}(k/2, 2s(1-s)) \geq \mathbf{B}(m_{V_0}, 2s(1-s)) + \mathbf{B}(m_{V_0,V_k}+m_{V_k}-k/2, 2s(1-s\cdot\max\{\gamma_{V_0,V_k}, \rho_{V_k}\})) + \mathbf{B}(k/2, 2s(1-s))$. Define $\widetilde{X} \sim \mathbf{B}(m_{V_0,V_k}+m_{V_k}-k/2, 2s(1-s\cdot\max\{\gamma_{V_0,V_k}, \rho_{V_k}\}))$ and $\widetilde{Y} \sim \mathbf{B}(m_{V_0,V_k}+m_{V_k}-k/2, 2s(1-s))$. Then, we have $\Pr(X \leq Y) \underset{\text{stochastically}}{=} \Pr(\widetilde{X} \leq \widetilde{Y}) \leq 2\exp(-\frac{1}{4}\frac{s^3(1-\max\{\gamma_{V_0,V_k},\rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k},\rho_{V_k}\}}(m_{V_0,V_k} + m_{V_k} - k/2)) \leq 2\exp(-2\ln n - 1) \leq \frac{1}{n^2}$. Since $\sum_{n>0}\frac{1}{n^2} = \frac{\pi^2}{6} < \infty$, it is a.a.s. that $\Pr(X \leq Y) \sim 0$ as $n \to \infty$, i.e., it is a.a.s. that $\Delta_{\sigma_0} < \Delta_{\sigma_k}$ given $\sigma_k \neq \sigma_0$.

(ii)Let $\mathbf{E}$ be the event that *there exists some $\sigma_k \neq \sigma_0$ and $\Delta_{\sigma_k} \leq \Delta_{\sigma_0}$*. Then, based on the union bound, we have $\Pr(\mathbf{E}) = \bigcup_{k=2}^{n-\Lambda}\Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) \leq \sum_{k=2}^{n-\Lambda} n^k\cdot\Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) \leq \sum_{k=2}^{n-\Lambda} n^k\cdot 2\exp(-(k+2)\ln n - \ln(2(n-\Lambda-1))) = \frac{1}{n^2}$. Consequently, it is a.a.s. that σ_0 is the only scheme inducing the least edge difference, i.e., it is a.a.s. that G^a is perfectly de-anonymizable. □

***Proof Sketch of Theorem 10*:** (i) As in Theorem 5, let V_c be the set of users that are perfectly de-anonymizable and $|V_c| = n-\epsilon n-\Lambda$. Furthermore, let $\mathbf{E}$ be the event that *there exists at least one incorrectly de-anonymizable user in V_c*. Then, we have $\Pr(\mathbf{E}) \leq \sum_{i\in V_c} Pr(i \text{ is incorrectly de-anonymized}) \leq \sum_{i\in V_c} 2\exp(-\frac{1}{4}\frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}}d_i) \leq \sum_{i\in V_c} 2\exp(-2\ln n - \ln(2(n-\epsilon n-\Lambda))) = \frac{1}{n^2}$. Consequently, it is a.a.s. that $\Pr(\mathbf{E}) \sim 0$ as $n \to \infty$, i.e. it is a.a.s. that G^a is $(1-\epsilon)$-de-anonymizable.

(ii) As in the proof of Theorem 5, we do not have to distinguish σ_k with σ_0 when $k \leq \epsilon n$. Let $\mathbf{E}$ be the event that *there exists some σ_k such that $\Delta_{\sigma_k} \leq \Delta_{\sigma_0}$ and $k > \epsilon n$*. Then, we have $\Pr(\mathbf{E}) \leq \sum_{k=\epsilon n+1}^{n-\Lambda} n^k\cdot\Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) \leq \sum_{k=\epsilon n+1}^{n-\Lambda} n^k\cdot 2\exp(-\frac{1}{4}\frac{s^3(1-\max\{\gamma_{V_0,V_k},\rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k},\rho_{V_k}\}}(m_{V_0,V_k}+m_{V_k}-k/2)) = \frac{1}{n^2}$. Consequently, it is a.a.s. that G^a is $(1-\epsilon)$-de-anonymizable as $n \to \infty$. □