Seed-Based De-Anonymizability Quantification of Social Networks

Shouling Ji, *Student Member, IEEE*, Weiqing Li, *Student Member, IEEE*, Neil Zhenqiang Gong, *Member, IEEE*, Prateek Mittal, and Raheem Beyah, *Senior Member, IEEE*

***Abstract*— In this paper, we implement the first comprehensive quantification of the perfect de-anonymizability and partial de-anonymizability of real-world social networks with seed information under general scenarios, which provides the theoretical foundation for the existing structure-based de-anonymization attacks and closes the gap between de-anonymization practice and theory. Based on our quantification, we conduct a large-scale evaluation of the de-anonymizability of 24 real-world social networks by quantitatively showing the conditions for perfectly and partially de-anonymizing a social network, how de-anonymizable a social network is, and how many users of a social network can be successfully de-anonymized. Furthermore, we show that both theoretically and experimentally, the overall structural information-based de-anonymization attack can be more powerful than the seed-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 paper. Our findings are expected to shed on research questions in the areas of structural data anonymization and de-anonymization and to help data owners evaluate their structural data vulnerability before data sharing and publishing.**

***Index Terms*— De-anonymization, social networks, quantification, evaluation.**

I. Introduction

With the development of information technology, social networks and services have become a permanent part of people's lives. The large amount of resulting social data is critical for academic research and has many important governmental and healthcare applications [3]. (1) *Academic Research:* It is well known that real-world social data are a valuable resource for academic researchers. There are annual events that are committed to sharing data, e.g., the KDD Cup events,[1] as well as many other academic events/institutions that regularly provide social data to the research community. This has enabled and/or enhanced research in many areas, including personalized advertising, secure routing, and Sybil detection. (2) *Government Applications:* Social data are frequently shared/transferred for government data mining tasks. For instance, customer understanding and international fraud detection can be achieved by leveraging the structure and pattern analysis of phone-call networks [32]. (3) *Business Applications:* Data sharing is now a common part of many companies' business models. For instance, as expressed in their privacy policies, Google, Facebook, and Twitter share their data with business partners for personalized precision advertising and targeted advising, under which cost savings and maximized advertising effectiveness can be achieved. (4) *Healthcare Applications:* Graph data are also shared for many healthcare-related applications. A typical application is the analysis of the propagation of infectious diseases, e.g., the flu, HIV, and Ebola.[2]

In contrast, social data increasingly contain the privacy information of users [3], [4], [21]. To protect users' privacy, the data owners (e.g., companies, government agencies, hospitals) usually anonymize their data before sharing, transferring, and/or publishing it. Generally, data anonymization techniques can be placed into four classes: *naive ID removal*, *k-anonymization (including randomly adding/deleting edges)* [12], [13], *differential privacy* [14]–[16], and other techniques [29], [31]. The naive ID removal method has been proven to be extremely vulnerable to state-of-the-art structure-based de-anonymization attacks [3], [4]. It also cannot be employed in k-anonymization to defend against structure-based de-aonymization attacks for real-world social networks due to its limitations, such as it not being scalable and richer information being available to adversaries. Differential privacy (and its variants) is designed to protect the privacy of data in *interactive queries* [14]. However, structure-based de-anonymization attacks are *non-interactive queries*. Thus, differential privacy cannot prevent such attacks in its current form (we discuss existing anonymization techniques in detail in the Related Work section).

Due to the vulnerability of existing anonymization schemes, the emerging *structure-based de-anonymization attacks* have been experimentally demonstrated to break the privacy of social networks effectively based only on the data's structural information, e.g., Narayanan and Shmatikov's

Manuscript received May 29, 2015; revised September 28, 2015 and December 16, 2015; accepted February 4, 2016. Date of publication February 11, 2016; date of current version April 5, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianying Zhou.

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

N. Z. Gong is with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: neilgong@iastate.edu).

P. Mittal is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08540 USA (e-mail: pmittal@princeton.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The file consists of proofs of some of the theorems in the paper and more experiment results. The material is 2.63 MB in size.

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/TIFS.2016.2529591

[1] http://www.sigkdd.org/kddcup/index.php

[2] http://www.andrew.cmu.edu/user/rkoganti/realistic.html; http://www.slideshare.net/jlcaut/ebola-hemoragic-fever-propagation-in-a-modern-city-using-sir-model

de-anonymization attack [3] and Srivatsa and Hicks' de-anonymization attack [4]. Although the de-anonymizability of social networks has been shown by experimental results (heuristic algorithms) in [3] and [4], the following are still open questions: *Why are social networks vulnerable to structure-based de-anonymization attacks? How de-anonymizable is a social network? How many users within a social network can be successfully de-anonymized?* Currently, there is some preliminary analysis on the de-anonymizability of social networks under the *Erdös-Rényi* (ER) random graph model or the *preferential attachment* model [7], [9], [10]. On the one hand, these existing analyses shed light on quantifying the de-anonymizability of social networks. On the other hand, however, all the existing analyses have some limitations, e.g., some did not consider the seed information, used an unrealistic network model, used unrealistic assumptions, or overlooked other more powerful structural information. These limitations prevent most existing analyses from being applicable to real-world social networks (the detailed discussion of the existing works and their limitations are presented in the Related Work section). To answer these open problems for general real-world social networks, we study the de-anonymizability of social networks in this paper. Specifically, our contributions can be summarized as follows.

1. To the best of our knowledge, we conduct the first seed-based theoretical quantification of the *perfect de-anonymizability* and *partial de-anonymizability* of social networks under the ER model as well as under general scenarios, where the social network can follow an arbitrary network model. Therefore, our quantification can be applied to real-world social networks and can quantitatively demonstrate the vulnerability of real-world social networks to existing structure-based de-anonymization attacks.

2. Based on our quantification, we implement a large-scale evaluation of the perfect and partial de-anonymizability of 24 real-world social networks. In our evaluation, we show the conditions for 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.

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

4. We discuss the implications of this paper and future work. Our quantification and evaluation enable understanding 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 research questions in the areas of structural data anonymization and de-anonymization. Furthermore, our quantification and evaluation are expected to attract the attention of data owners and help them develop more proper privacy protection policies.

Differences Between This Work and Previous Works. A preliminary version [1] of our paper was published at NDSS 2015. In this journal version, we added more than 30% new content, including both theoretical analysis and experimental evaluation. Specifically, we emphasize the following: (1) In the conference version, we only provided the proof sketches of Theorems 5, 6, 8, and 10 due to space limitations. In this version, we provide the complete proofs of the four theorems along with explanations. (2) We stated our other main theoretical conclusions, Theorems 1, 2, and 3, in the conference version without justification. In this version, we formally prove them, which completes our perfect and partial de-anonymizability quantification and makes it easy to follow. (3) In this version, we added a new section to evaluate the condition for n for the $(1-\epsilon)$-de-anonymizability of the 24 social networks. We also analyzed the new evaluation results in detail. (4) Compared to the conference version, we also provided more explanations to make the proposed technique more understandable.

The rest of this paper is organized as follows. In Section II, we describe the system model, assumptions, and problem definition. The preliminary quantification under the ER model is implemented in Section III. We quantify the perfect and partial de-anonymizability of social networks under general scenarios in Section IV. In Section V, we evaluate the de-anonymizability of 24 real-world social networks. Finally, the paper is concluded in Section VII. We summarize the related work with remarks and provide more evaluation results in the *Supplementary File*.

II. System Model, Assumptions, and Definitions

In this section, we introduce the system model and related assumptions and definitions. For the sake of readability, we have summarized the frequently used acronyms and symbols in Table I.

A. Data Model

In our quantification and evaluation, we use the same graph model used in in [2]–[10] 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 $E^a = \{e^a_{i,j} | i, j \in V^a$, is a social tie that exists between i and $j\}$. To de-anonymize G^a, we use an auxiliary social network that has overlapping users with G^a and can be obtained using multiple methods, e.g., data aggregation, data mining, collaborative information systems, knowledge/data

TABLE I
SUMMARY OF THE NOTATIONS

notation	definition
$G^a = (V^a, E^a)$	anonymized graph
$G^u = (V^u, E^u)$	auxiliary graph
i, j	nodes/users
n	number of users
$e^a_{i,j}, e^u_{i,j}$	social tie (links/edges)
d^a_i, d^u_i	degree of i
s_a, s_u, s	graph sampling probabilities
σ	a de-anonymization scheme
σ_0	the perfect de-anonymization
σ_k	a de-anonymization scheme with k errors
$\mathcal{S}$	seed mappings
$\Lambda = \lvert\mathcal{S}\rvert$	the cardinality of $\mathcal{S}$
$\Delta_{\sigma:(i,j)}$	edge difference induced by $(i,j) \in \sigma$
Δ_σ	edge difference of σ
$G(n,p)$	ER random graph with parameters n and s
ϵ	tolerated de-anonymization error
m	number of edges
ρ, ρ_U	graph density
$\gamma_{U,W}$	graph connectivity

brokers [2]–[10], [19], [20].[3] 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 $E^u = \{e^u_{i,j} | i, j \in V^u$, is a social tie that exists between i and $j\}$. For theoretical quantification, without describing too many of the mathematical details, we assume that both G^a and G^u are undirected graphs.[4] Furthermore, because our quantification and evaluation are based on the graph model, our work can be potentially applied to other kinds of data that can be modeled using 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.

B. Graph Sampling

To make the quantification mathematically tractable, we employ the same assumptions on G^a and G^u as those used in [7], [9], and [10]. First, $V^a = V^u = \{1, 2, \cdots, n\}$ [7], [9], [10]. 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, i.e., adding dissimilar users to each graph with degree zero to make V^a and V^u mathematically equivalent. Note that this is a mathematical assumption that does not limit the generality of this work. Our quantification is also valid for 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 true relationships among users in V* [7], [9], [10]. In particular, we assume that 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 because people are usually involved in multiple social networks and G^a and G^u are some particular social networks in which the users in V are involved. For instance, G^a could be LinkedIn (a professional social network of V), and G^u could be Facebook (a social network of V based on friendships).

C. 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 $\sigma(i) \in V^u$. Because $V^a = V^u$, for simplicity, we define a *successful de-anonymization* of $i \in V^a$ achieved under σ if $i = \sigma(i)$. In addition, we use σ_0 to denote the *perfect de-anonymization*, i.e., $\sigma_0 = \{(i,i) | i = 1, 2, \cdots, n\}$ (all the users in G^a are correctly de-anonymized), 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, $k \in [2, n]$. In the rest of this paper, we say that $i \in V^a$ is perfectly de-anonymizable if i can be correctly de-anonymized, and V^a is perfectly de-anonymizable if all the users in V^a can be correctly de-anonymized.

Most existing de-anonymization algorithms (e.g., [2]–[4]) consist of two phases: *seed identification phase* that identifies some *seed mapping information* from V^a to V^u, and *mapping propagation phase* that 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 [2]–[4], we assume that we have identified a *seed mapping set* from V^a to V^u by some technique (e.g., the methods in [2]–[4]), 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. We now have to quantify the de-anonymizability of a social network G^a given $\mathcal{S}$, G^u and the existence of G, s_a, and s_u.

To make the quantification easy to follow and the conclusions succinct, we further assume $s_a = s_u = s$, i.e., we assume G^a and G^u are two instances of G with the same sampling probability. Note that this assumption does not change our analysis in terms of any material detail. All our quantification

[3]For the detailed means of obtaining the auxiliary social network, please refer to the discussion in [3] and [8]. In particular, with the emergence of data brokers, many auxiliary data can be easily obtained at an affordable cost.

[4]In reality, many graph data carry direction information, i.e., they are directed graphs. Furthermore, some de-anonymization attacks are designed to utilize the direction information to improve the de-anonymization performance, e.g., [3]. In this paper, we do not take into account the direction information mainly because we want to make our quantification sufficiently general. Although our quantifications are based on the undirected graph model, they can be extended to directed graphs directly by overlooking the direction information on the edges.

Nevertheless, when applying our quantifications to directed graphs, overlooking the direction information may lead to inaccurate de-anonymizability quantification (potentially underestimating the de-anonymizability of the data). The impact of direction information on the de-anonymizability of graph data itself is an interesting research topic and requires a proper model to characterize the direction information, elegant quantification techniques, and dedicated research. We will undertake this research as one of our future research directions.

results can be extended to the case $s_a \neq s_u$, only with more complex expressions.

D. Measuring σ

Given G^a, G^u, and a de-anonymization scheme σ, we measure σ as 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 let $\sigma(E^a_i(A)) = \{\sigma(e^a_{i,v}) | e^a_{i,v} \in E^a_i(A)\}$ ($\sigma(e^u_{i,j})$, $E^u_i(A)$, 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

$$\Delta_{\sigma:(i,j)} = |\sigma(E^a_i) \setminus E^u_j| + |\sigma^{-1}(E^u_j) \setminus E^a_i|, \tag{1}$$

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

$$\Delta_\sigma = \sum_{(i,j)\in\sigma} \Delta_{\sigma:(i,j)}, \tag{2}$$

which indicates the edge difference between G^a and G^u under σ. Intuitively, because 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 III and IV).

Similar to $\Delta_{\sigma:(i,j)}$ and Δ_σ, we define $\Delta_{\sigma:(i,j)}(\mathcal{S})$, which measures the edge difference of a mapping (i, j) with respect to $\mathcal{S}$:

$$\begin{aligned}\Delta_{\sigma:(i,j)}(\mathcal{S}) = {} & |\sigma(E^a_i(\mathcal{S}^a) \setminus E^u_j(\mathcal{S}^u)| \\ & + |\sigma^{-1}(E^u_j(\mathcal{S}^u) \setminus E^a_i(\mathcal{S}^a)|,\end{aligned} \tag{3}$$

and $\Delta_\sigma(\mathcal{S})$, which measures the edge difference of a de-anonymization scheme σ with respect to $\mathcal{S}$:

$$\Delta_\sigma(\mathcal{S}) = \sum_{(i,j)\in\sigma} \Delta_{\sigma:(i,j)}(\mathcal{S}). \tag{4}$$

III. Quantification Under the Erdös-Rényi Model

In this section, we quantify the de-anonymizability of G^a with $\mathcal{S}$, G^u, G, and s under the Erdös-Rényi (ER) model, i.e., we assume that $G(V, E)$ is a *random graph* generated from the ER model $G(n, p)$, where n is the number of nodes in the graph and p specifies the probability of an edge existing between any pair of nodes. Although real-world social networks rarely satisfy the ER model [21], the analysis in this section can shed the light on the quantification under general scenarios (Section IV). For the sake of readability, we provide all the proofs in Section II of *Supplementary File*.

A. $\mathcal{S}$-Based Quantification

We first quantify the de-anonymizability of G^a based only on the seed information $\mathcal{S}$. For the de-anonymization scheme σ, we assume that 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 $\Delta_{\sigma:(i,\sigma(i))}(\mathcal{S})$.[5]

[5]Because our focus is on quantifying the de-anonymizability of G^a, we do not consider the actual de-anonymization algorithms. Specifically, we aim to provide the theoretical foundation for the workability of structure-based de-anonymization attacks, e.g., [2]–[4].

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 can be stated* asymptotically almost surely (a.a.s.)[6] *that* $\forall i \in V^a \setminus \mathcal{S}^a$*, i is perfectly de-anonymizable (i can be successfully de-anonymized).*

In Theorem 1, we quantify the condition for p, s, and $\mathcal{S}$ on 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 for p, s, and $\mathcal{S}$ such that all the users in $V^a \setminus \mathcal{S}^a$ are perfectly de-anonymizable.

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 can be stated a.a.s. that all the users in* $V^a \setminus \mathcal{S}^a$ *are perfectly de-anonymizable.*

B. Sophisticated Quantification: Considering More Structure Information

In the previous subsection, we quantified the de-anonymizability of G^a based only on the seed knowledge. Actually, besides the edges in $E^a_i(\mathcal{S})/E^u_i(\mathcal{S})$, all the edges in E^a_i/E^u_i can provide structure information that can be used for de-anonymization. In this subsection, we 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 only scheme inducing the least edge difference. Basically, the first part of Theorem 3 can be proven using a technique similar to that used in [7]. Here, we obtain a tighter bound by applying more elegant quantification techniques.

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 can be said a.a.s. that* $\Delta_{\sigma_0} < \Delta_{\sigma_k}$ ($k \in [2, n]$)*, i.e., it can be said a.a.s. that the perfect de-anonymization scheme σ_0 induces a smaller number of edge differences 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 can be said a.a.s. that the perfect de-anonymization scheme σ_0 induces the least edge difference compared to all the other de-anonymization schemes, i.e., it can be said a.a.s. that σ_0 is the only scheme inducing the least edge difference.*

Theorem 3 has a very 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 can be said a.a.s. that the perfect de-anonymization scheme σ_0 induces the least edge difference compared to all the other de-anonymization schemes, i.e., it can be said a.a.s. that σ_0 is the only scheme inducing the least edge difference.*

Based on Theorems 2 and 3 and Corollary 1, it is straightforward to obtain a more accurate (tighter) bound on the

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

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

We now study the structural condition for 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 to be $(1-\epsilon)$-*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 $(1-\epsilon)$-de-anonymizable with or without seed information in Theorem 5, i.e., the condition that at most ϵn incorrect de-anonymizations are allowable.

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.*

IV. Quantification in General Scenarios

Although the ER model is suitable for elegant theoretical analysis of the de-anonymizability of social networks, the fact is that it is extremely rare, if not impossible, to observe real-world social networks actually following the ER model [21]. Nevertheless, the ER model can shed light on the theoretical quantification of the de-anonymizability of social networks under general scenarios.

In this section, we quantify the de-anonymizability of G^a under general scenarios, i.e., unlike in Section III, we assume $G(V,E)$ could now be some graph following an arbitrary network model. To accelerate the quantification, we make the following definitions. 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 ($U\cap W=\emptyset$), and let $E_{U,W}=\{e_{i,j}\in E|i\in U, j\in W\}$ 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, we assume that the seed mapping set $\mathcal{S}$ is randomly identified, which implies that each user in V is selected with a probability $q=\frac{\Lambda}{n}$. We denote the seed users in V as a set S for convenience, i.e., $S=\mathcal{S}^a=\mathcal{S}^u$. We denote the other users by set $A=V\setminus S$.

For the sake of readability, we provide all the proofs in Section II of *Supplementary File*.

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.

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 can be said a.a.s. that* $\forall i\in A$, i *is perfectly de-anonymizable.*

In Theorem 6, the condition where a user is perfectly de-anonymized is quantified. We further quantify the condition for perfectly de-anonymizing 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 can be said a.a.s. that* G^a *is perfectly de-anonymizable.*

B. Sophisticated Quantification: Considering More Structure Information

In the previous subsection, the perfect de-anonymizability of social networks is quantified under general scenarios based on $\mathcal{S}$. As we discussed in Section III, 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 for improving the de-anonymization performance (as shown in Theorem 3). Similar to quantification under the ER model, we quantify the de-anonymizability of social networks by considering the overall structure information in Theorem 8.

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 can be said a.a.s. that* $\Delta_{\sigma_0}<\Delta_{\sigma_k}$ $(k\in[2,n])$, *i.e., it can be said a.a.s. that the perfect de-anonymization scheme* σ_0 *induces a lower 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 can be said 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 to Theorem 3, Theorem 8 also implies that a large-scale social network is perfectly de-anonymizable without seed information under 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 can be said 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 and 8 and Corollary 2, the following conclusion is straightforward.

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 can be said 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 under general scenarios, where ϵn ($\epsilon\in[0,1-\frac{\Lambda}{n}]$) users are now allowed to be incorrectly de-anonymized. We demonstrate the quantification in Theorem 10.

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.*

TABLE II
DATASET STATISTICS

Name	n	m	ρ	$\overline{d}$	$p(1)$	$p(5)$	$p(10)$
Hyves	1,402,673	2,777,419	2.82E-06	3.96	56.76%	88.74%	91.80%
Douban	154,908	327,162	2.73E-05	4.22	66.57%	90.81%	93.86%
Friendster	5,689,498	14,067,887	8.69E-07	4.95	60.19%	91.27%	95.86%
YouTube	1,138,499	2,990,443	4.61E-06	5.25	53.16%	85.53%	92.78%
Flixster	2,523,386	7,918,801	2.49E-06	6.28	59.49%	87.26%	92.86%
Last.fm	1,191,812	4,519,340	6.36E-06	7.58	47.27%	81.62%	89.54%
FB-NO-wall	45,813	183,412	1.75E-04	8.01	24.18%	60.91%	77.42%
Gowalla	196,591	950,327	4.92E-05	9.70	25.20%	64.50%	79.90%
Foursquare	639,014	3,214,986	1.57E-05	10.06	51.10%	79.11%	83.21%
Enron	33,696	180,811	3.19E-04	10.73	28.09%	67.86%	82.88%
Skitter	1,694,616	11,094,209	7.73E-06	13.09	12.80%	55.41%	76.21%
Slashdot	82,168	582,533	1.73E-04	14.18	2.19%	64.78%	78.30%
Digg	771,229	5,907,413	1.99E-05	15.32	45.64%	77.31%	85.97%
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90	20.99%	50.53%	64.88%
HepPh	11,204	117,649	1.87E-03	21.00	9.95%	49.99%	66.45%
AstroPh	17,903	197,031	1.23E-03	22.01	5.34%	33.69%	50.66%
FB-NO-links	63,731	817,090	4.02E-04	25.64	12.71%	36.11%	50.02%
Pokec	1,632,803	22,301,964	1.67E-05	27.32	10.04%	30.66%	44.48%
BlogCatalog	97,884	1,668,647	3.48E-04	34.10	28.24%	59.59%	71.45%
Google+	4,692,671	90,751,480	8.24E-06	38.68	5.44%	27.33%	46.37%
Livemocha	104,103	2,193,083	4.05E-04	42.13	6.56%	27.56%	44.02%
Twitter	456,293	12,508,272	1.20E-04	54.83	5.30%	19.76%	34.50%
Orkut	3,072,441	117,185,083	2.48E-05	76.28	2.21%	7.28%	13.35%
Flickr	80,513	5,899,882	1.82E-03	146.56	0.00%	11.63%	20.58%

V. LARGE-SCALE EVALUATION

A. Datasets and Set Up

We use 24 various real-world social datasets that are mainly from the *Stanford Large Network Dataset Collection* [22], *ASU Social Computing Data Repository* [23], and other sources [24], [25]. The datasets are shown in Table II 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 $p(k)$ $(k = 1, 5, 10)$ is the percentage of users with degree less than or equal to k. We further briefly introduce the datasets in Section III of *Supplementary File*. The detailed descriptions can be found in the corresponding references.

For each dataset, we use the raw data except when 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. Because the *direction information* can be used to improve the effectiveness of de-anonymization attacks [3], it is possible that our quantification and evaluation can be improved if we have more knowledge, e.g., the direction information. In future work, we aim to quantify the de-anonymizability of directed social networks.

To generate the anonymized and auxiliary graphs, we follow the data sampling approach discussed in Section II, i.e., we construct G^a and G^u from the raw data using the sampling probabilities s_a and s_u, respectively. Here, for simplicity, we set $s_a = s_u = s$. After constructing G^a and G^u, the seed mappings are chosen randomly from them (note that seed mappings are some pre-known user mappings between G^a and G^u), which implies that the high-degree users are not given preference, as in [9] and [10], although they may be more helpful as seed mappings. Consequently, our evaluation results represent the general results of our quantification. Each group of evaluations is repeated 50 times, and the results are the averages of these 50 runs.

We quantify the de-anonymizability of a social network using seed information and using the overall structural information. 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 the overall structural information, respectively. *Not specifying the suffix or the particular context implies using the overall structural information by default.* Due to space limitations (and for the sake of readability), we evaluate the condition for the perfect de-anonymizability of the datasets in Table II in *Supplementary File*.

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

More details on the evaluation of $(1-\epsilon)$-de-anonymizability can be found in *Supplementary File*.

1) Evaluation of $(1-\epsilon)$: 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*.

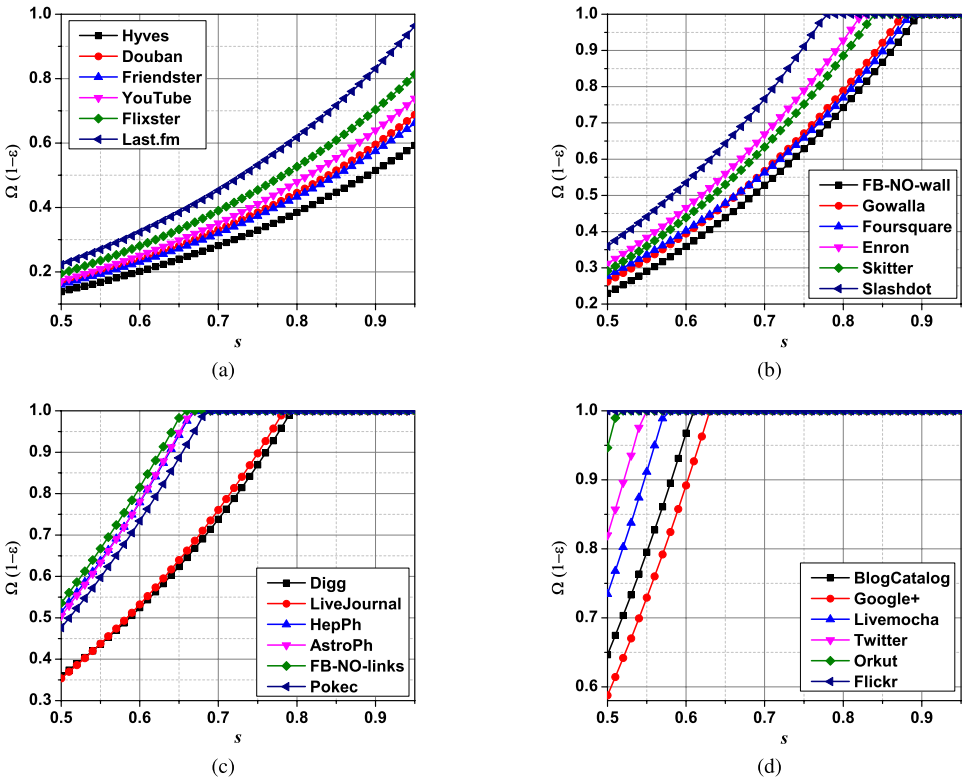


Fig. 1. $(1-\epsilon)$-de-anonymization: $\Omega(1-\epsilon)$ vs. s. Default setting: $\Lambda = 0.05n$ (5% users are seeds).

When all the structural information (including seed mappings) is 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. 1 with different s. From Fig. 1, we have the following observations.

- 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 de-anonymizable based on the overall structural information. Consequently, the obtained quantitative results confirmed the success of existing heuristic algorithms [3], [4]. This is also consistent with our quantification of $(1-\epsilon)$-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 de-anonymizable because they carry enough structural information.
- When s increases, $\Omega(1-\epsilon)$ also increases, i.e., 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 for this is similar to the explanation in the previous subsection: a larger s implies that more common edges are 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 correct user de-anonymization induces a lower edge difference (de-anonymization error).
- When s is increased above some value, several social networks can be asymptotically perfectly de-anonymizable ($\Theta(n)$ 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. The reason behind this is the same that for the previous observation: a larger s implies more structural similarity and a more de-anonymizable social network.
- The social networks with a higher average degree $\overline{d}$ are more de-anonymizable, e.g., when $s = 0.6$, 53.23% LiveJournal users ($\overline{d} = 17.9$) are perfectly de-anonymizable, while 73.38% Pokec users ($\overline{d} = 27.32$) are perfectly 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 more likely induces a lower edge difference.

We now study the $(1-\epsilon)$-de-anonymizability of the 24 social networks when we fix the network density, s, and Λ/n while varying n. The results are shown in Fig. 2. From Fig. 2, we have the following observations.

- When n increases, the percentage of de-anonymizable users of each social network also increases for both seed-based de-anonymization 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 the 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

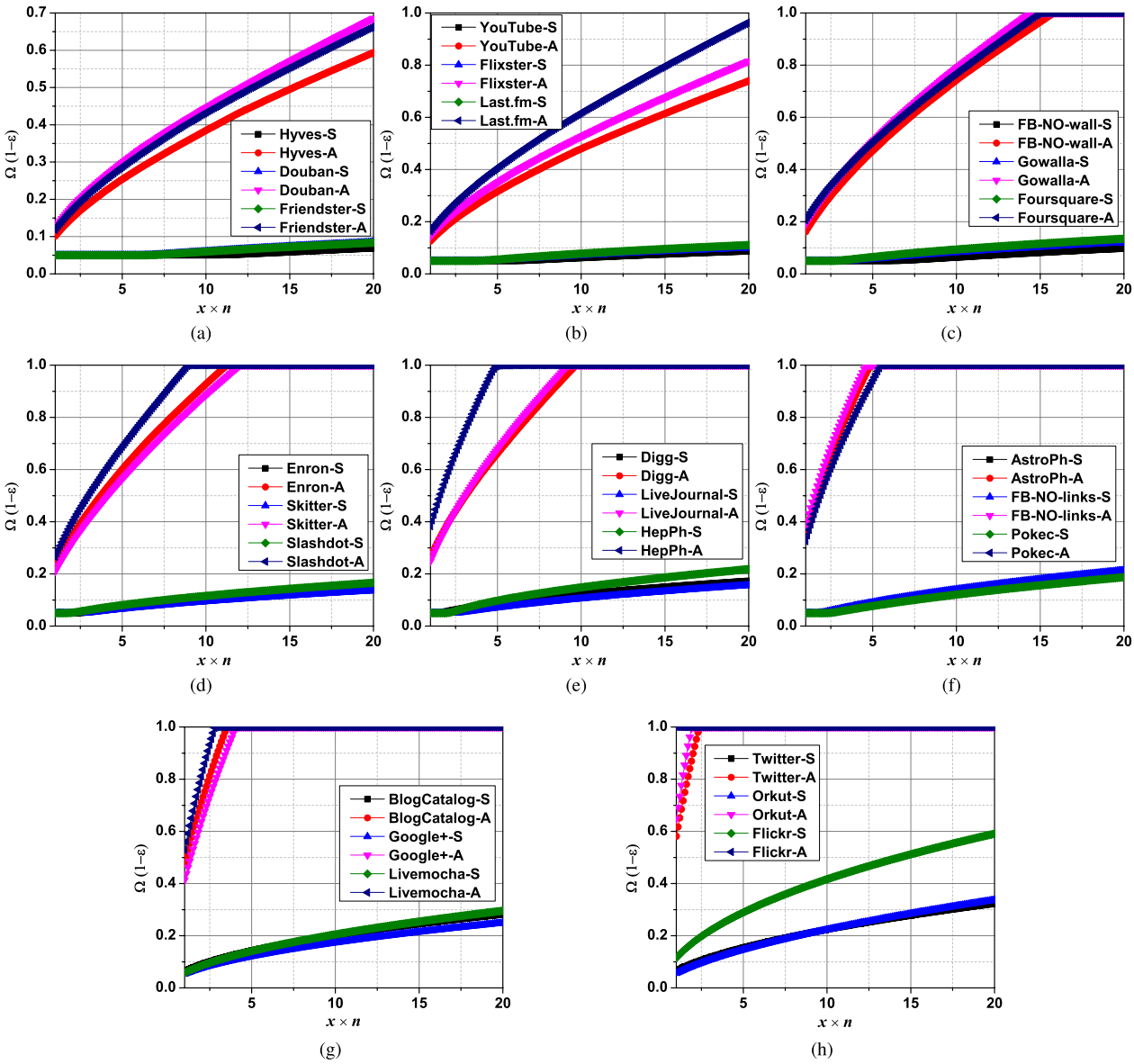


Fig. 2. $(1-\epsilon)$-de-anonymization: $\Omega(1-\epsilon)$ vs. n. Default setting: $s = 0.8$ and $\Lambda/n = 0.05$.

our quantification. A large n implies richer structural information when ρ is fixed. Hence, more users are de-anonymizable.

- As expected, the overall structural information is more powerful in de-anonymizing social networks. This is also consistent with our quantification. Because more structural information is considered, the probability that correct de-anonymization induces more edge differences than incorrect de-anonymization is decreased. Consequently, "*-A" de-anonymizes more users than "*-S".
- As validated before, the graph density also has a positive impact on $\Omega(1-\epsilon)$, i.e., a social network with a high graph density is more de-anonymizable, as a higher 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 if 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. 3. From Fig. 3, we have the following observations.

- 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, as more seed mappings imply that more knowledge is available to improve the de-anonymization accuracy, which can also be seen from our quantification.
- 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 $\overline{d}$ may be more de-anonymizable than a social network with higher ρ or $\overline{d}$ in some cases, e.g., BlogCatalog has a smaller $\overline{d}$ but larger ρ than Google+, and Orkut has a smaller ρ but larger $\overline{d}$ than BlogCatalog. This is because the seed mappings 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.

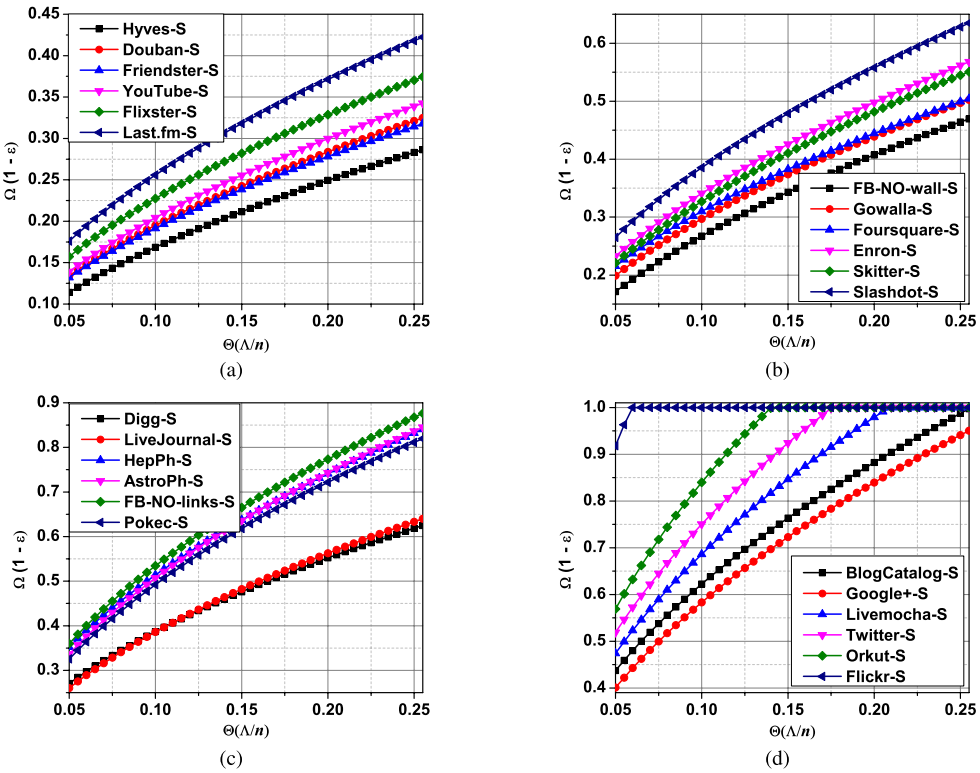


Fig. 3. $(1-\epsilon)$-de-anonymization: $\Omega(1-\epsilon)$ vs. Λ. Default setting: $s = 0.8$.

However, it is difficult to determine which one will dominate the de-anonymizability. Generally speaking, the richer is the structural information, i.e., the higher ρ and $\overline{d}$ are, the more de-anonymizable the social network is.

2) Evaluation of Λ: In this subsection, we evaluate the condition for Λ in $(1-\epsilon)$-de-anonymization. When $\epsilon = 0.4$, i.e., up to a 40% user de-anonymization error is tolerable, the condition for Λ to perfectly de-anonymize at least $1-\epsilon = 60\%$ users of each social network under different settings of s is shown in Fig. 4. From Fig. 4, we can observe the following:

- 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 for 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, a smaller number of common edges are shared by G^a and G^u. Consequently, all the anonymized users tend to be involved as seeds to achieve perfect de-anonymizability.
- For seed-based de-anonymization, when s is above some threshold value, $\Theta(\Lambda/n)$ decreases with increases in s (a smaller number of seed mappings is 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 can be said 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 (i) 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; (ii) 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 for Λ to make each social network $(1-\epsilon)$-de-anonymizable under different ϵ is shown in Fig. 5. From Fig. 5, we can see the following:

- In seed-based de-anonymization, to make social networks with a 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. 5 (a)-(d) have $\overline{d} < 15$, and the condition for Λ to make them $(1-\epsilon)$-de-anonymizable is $\Theta(\Lambda/n) \sim 1-\epsilon$. The reason for this is that a low $\overline{d}$ implies a lower number of 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. 5 (e)-(h), a lower number of seed mappings need 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% users are to serve as seeds.
- 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 for this is that when the overall structural information (including seed mappings) is considered and $s = 0.8$, the correct de-anonymization induces

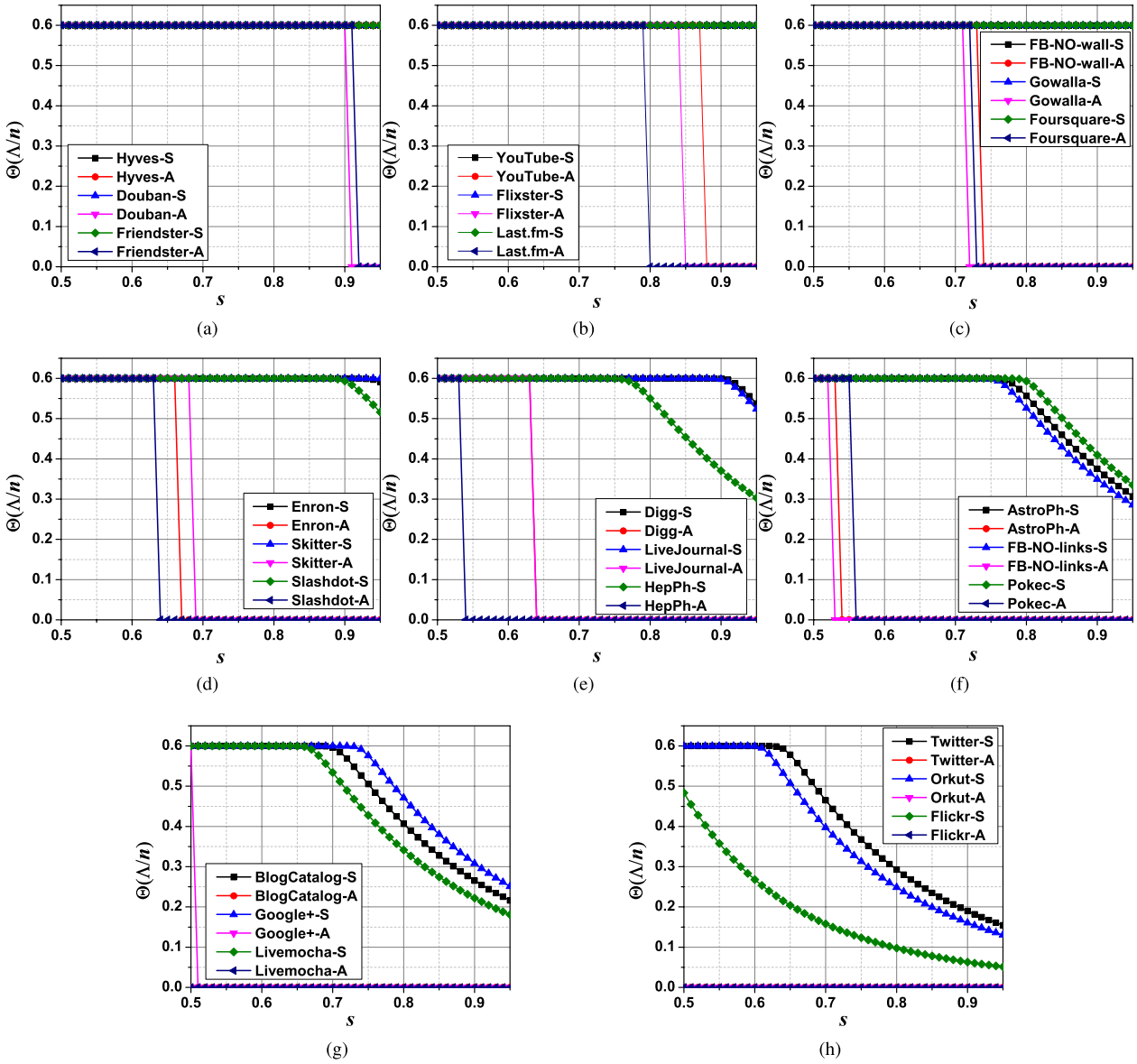


Fig. 4. $(1-\epsilon)$-de-anonymization: Λ vs. s. Default setting: $\epsilon = 0.4$.

the lowest number of edge differences with higher probability than in seed-based de-anonymization, which is consistent with our quantification. Again, the results also confirm that the overall structural information-based de-anonymization is more effective.

We now evaluate the condition for Λ when the network size changes while other network properties are fixed. The results are shown in Fig. 6. From Fig. 6, we make the following observations.

- 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. 6 (a)-(d), it is necessary to have $\Theta(\Lambda/n) \sim 1-\epsilon$ in seed-based de-anonymization. The reason for this is also similar to that presented in the earlier analysis. A small $\overline{d}$ implies a lower number of 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 continues to increase, a lower number of seed mappings are needed for social networks with high $\overline{d}$ (social networks in Fig. 6 (e)-(h)) to be $(1-\epsilon)$-de-anonymizable. The reason for this is also similar to that presented in the earlier analysis.
- 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 provides the foundation for de-anonymization attack without seed information.

VI. Discussion and Future Work

A. Discussion

In this paper, to the best of our knowledge, we conduct the first comprehensive theoretical quantification of the de-anonymizability of social networks (e.g., Facebook,

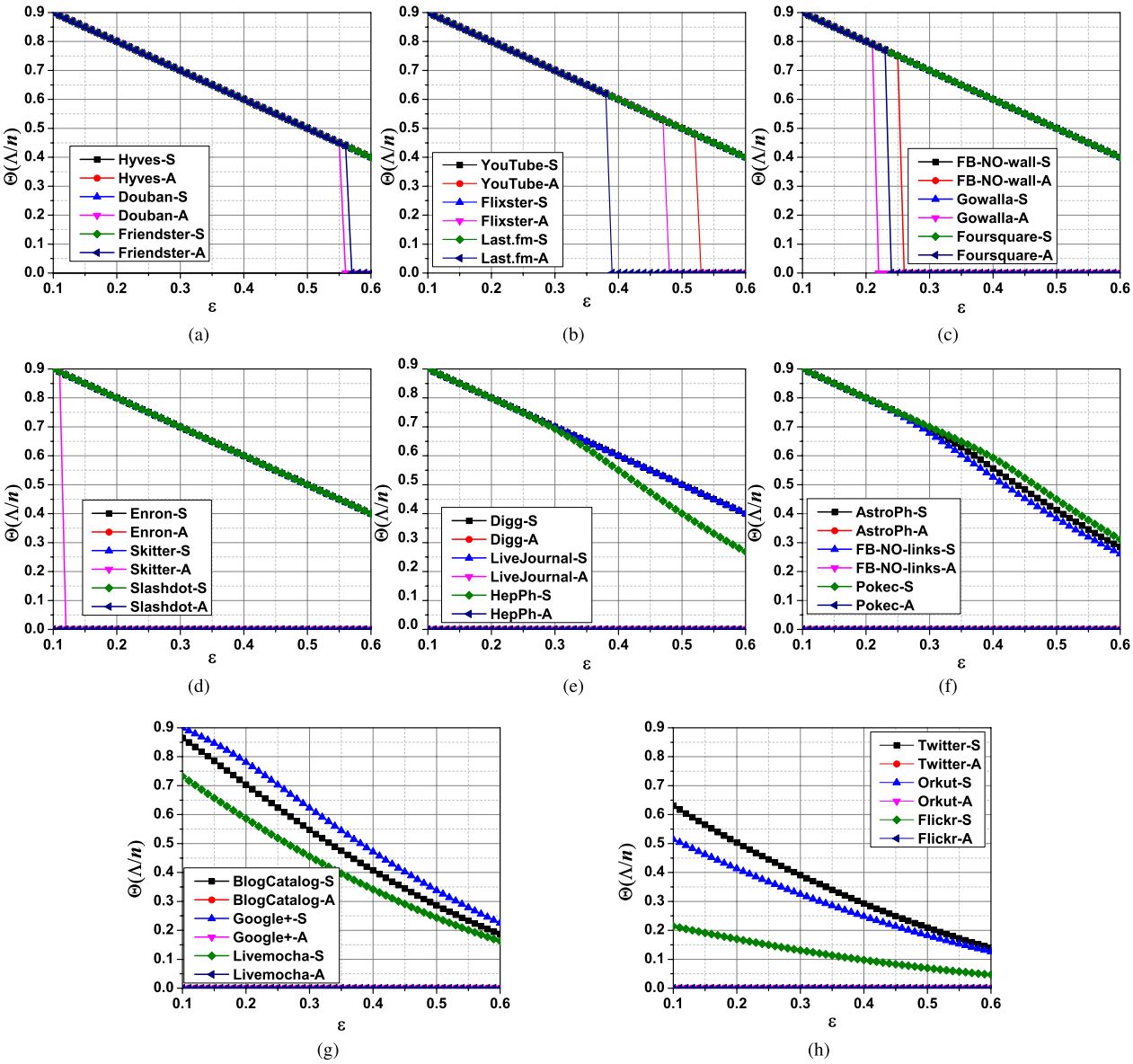


Fig. 5. $(1-\epsilon)$-de-anonymization: Λ vs. ϵ. Default setting: $s = 0.8$.

Google+, Twitter) under both the theoretical ER model and the general arbitrary network model. The most meaningful significance of our quantification is that it provides the theoretical foundation for the existing *de-anonymization attacks with available seed information* [3], [4], i.e., for the first time, we theoretically demonstrate that structure-based de-anonymization attacks are sound. This closes the gap between existing heuristic de-anonymization algorithms (e.g., Backstrom et al.'s de-anonymization attack [2], Narayanan and Shmatikov's de-anonymization attack [3], Srivatsa and Hicks' de-anonymization attack [4]) and their theoretical foundation.

Further limitations of this paper are summarized as follows. 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. Some other global graph properties are also helpful in improving structure-based de-anonymization attacks, e.g., the *betweenness centrality* and the *closeness centrality* of a user and the distance from a user to all the other users. Although we believe these graph properties can be used in improving de-anonymization attacks, it is difficult to include them in the theoretical quantification. 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 is very difficult, if not impossible, to quantify the change in 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 for perfectly or partially de-anonymizing a social network.

In this paper, we focus on closing the gap between existing de-anonymization practice (i.e., heuristic algorithms)

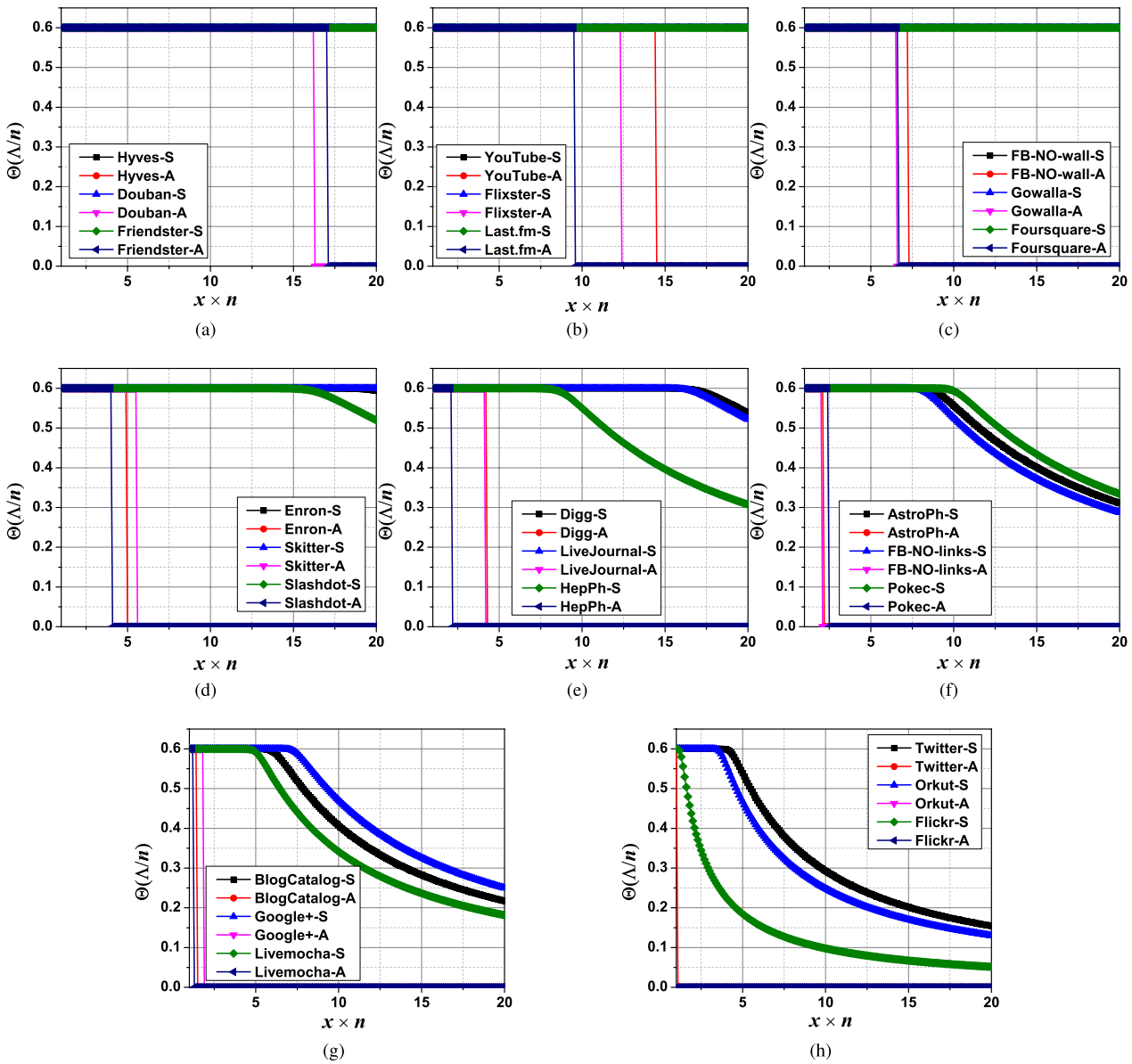


Fig. 6. $(1-\epsilon)$-de-anonymization: Λ vs. n. Default setting: $s = 0.8$ and $\epsilon = 0.4$.

and its 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. This is still an important open problem because we have an increasing amount of social data. We believe our quantification and evaluation in this paper can shed light on future research questions in this area by providing the theoretical foundation for structure-based de-anonymization attacks and their effectiveness in attacking real-world social networks. Furthermore, our quantification and evaluation are expected to attract the attention of data owners and help them develop more proper policies to protect social data.

B. Future Work

Our future work will take the following directions: (i) We expect to develop a new mathematical model under which we can theoretically analyze the change in users' global topological properties (e.g., betweenness/closeness centrality, the distance to seed users). Then, we can quantify the de-anonymizability of social networks more accurately. (ii) In our quantification, we simply assume that G^a and G^u are two sampling versions of an underlying conceptual graph G. In the future, we propose to remove this assumption by studying more practical and general models to characterize the structural correlation between the anonymized graph and the auxiliary graph. (iii) In our quantification, we do not explicitly involve the noise level because we do not involve a specific *noise description model* (actually, we currently do not have proper schemes to add noise with usability preservation, to the best of our knowledge). In the future, we propose to quantify the de-anonymizability of social networks by involving a function describing the existing noise. (iv) Our quantifications are conducted based on undirected graphs. Although they can be applied to directed graphs, the de-anonymizability of a graph

may be underestimated (as shown in [3], direction information can be used to improve the de-anonymization performance). Therefore, we plan to quantify the de-anonymizability of directed graphs. (v) As pointed out previously, we do not have effective data anonymization techniques for 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* that can examine the data de-anonymizability, anonymize data properly with usability preservation, and publish data securely. We also propose to develop new social data protection policies for data owners (e.g., companies, government agencies, hospitals.).

VII. Conclusion

In this paper, we study the de-anonymizability of social networks based only on their structural information. First, we quantify the *perfect de-anonymizability* and $(1-\epsilon)$-*de-anonymizability* of social networks with seed information under the mathematical ER model. Subsequently, we extend our quantification to general scenarios, where a social network can follow an arbitrary network. Third, based on our quantification, we conduct a large-scale evaluation of the de-anonymizability of 24 real-world social networks. Finally, we discuss the implications of this work. Our findings are expected to shed light on research questions in structural data anonymization and de-anonymization and help data owners evaluate their data vulnerability before data sharing/publishing.

References

[1] S. Ji, W. Li, N. Z. Gong, P. Mittal, and R. Beyah, "On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge," in *Proc. NDSS*, Feb. 2015, pp. 1–15.

[2] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou R3579x? Anonymized social networks, hidden patterns, and structural steganography," in *Proc. WWW*, 2007, pp. 181–190.

[3] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proc. S&P*, May 2009, pp. 173–187.

[4] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social network as a side-channel," in *Proc. 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. ISC*, 2014, pp. 237–254.

[6] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, "Community-enhanced de-anonymization of online social networks," in *Proc. CCS*, 2014, pp. 537–548.

[7] P. Pedarsani and M. Grossglauser, "On the privacy of anonymized networks," in *Proc. KDD*, 2011, pp. 1235–1243.

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

[9] L. Yartseva and M. Grossglauser, "On the performance of percolation graph matching," in *Proc. COSN*, 2013, pp. 119–130.

[10] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," in *Proc. PVLDB*, 2014, pp. 377–388.

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

[12] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting structural re-identification in anonymized social networks," in *Proc. VLDB*, 2008, pp. 102–114.

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

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

[15] N. Li, W. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy," in *Proc. ASIACCS*, 2012, pp. 32–33.

[16] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang, "Membership privacy: A unifying framework for privacy definitions," in *Proc. CCS*, 2013, pp. 889–900.

[17] A. Sala, X. Zhao, C. Wilson, H. Zheng, and B. Y. Zhao, "Sharing graphs using differentially private graph models," in *Proc. IMC*, 2011, pp. 81–98.

[18] A. Meyerson and R. Williams, "On the complexity of optimal K-anonymity," in *Proc. PODS*, 2004, pp. 223–228.

[19] C. Shah, R. Capra, and P. Hansen, "Collaborative information seeking [Guest editors' introduction]," *Computer*, vol. 47, no. 3, pp. 22–25, Mar. 2014.

[20] 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.

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

[22] *Stanford Large Network Dataset Collection*. [Online]. Available: http://snap.stanford.edu/data/index.html, accessed Mar. 23, 2010.

[23] *Social Computing Data Repository at ASU*. [Online]. Available: http://socialcomputing.asu.edu/pages/datasets

[24] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," in *Proc. WOSN*, 2009, pp. 37–42.

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

[26] D. Goodin. *Poorly Anonymized Logs Reveal NYC Cab Drivers' Detailed Whereabouts*. [Online]. Available: http://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/, accessed Jun. 23, 2014.

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

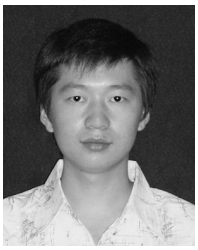
[28] H. Yu, C. Shi, M. Kaminsky, P. B. Gibbons, and F. Xiao, "DSybil: Optimal sybil-resistance for recommendation systems," in *Proc. S&P*, May 2009, pp. 283–298.

[29] F. Beato, M. Conti, and B. Preneel, "Friend in the middle (FiM): Tackling de-anonymization in social networks," in *Proc. SECSOC*, Mar. 2013, pp. 279–284.

[30] G. G. Gulyás and S. Imre, "Measuring importance of seeding for structural de-anonymization attacks in social networks," in *Proc. SECSOC*, Mar. 2014, pp. 610–615.

[31] G. G. Gulyás and S. Imre, "Using identity separation against de-anonymization of social networks," *Trans. Data Privacy*, vol. 8, no. 2, pp. 113–140, 2015.

[32] G. J. Wills, "NicheWorks—Interactive visualization of very large graphs," *J. Comput. Graph. Statist.*, vol. 8, no. 2, pp. 190–212, 1999.



Shouling Ji (S'10) received 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 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, differential privacy, password security, and machine learning security and privacy. He also has interests in graph theory and algorithms and wireless networks. He is a Student Member of the Association for Computing Machinery and was the Membership Chair of the IEEE Student Branch at Georgia State (2012–2013).

Weiqing Li (S'14) 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 with the School of Electrical and Computer Engineering. His research interests include big data privacy and network security. He is a Student Member of the Association for Computing Machinery.

Neil Zhenqiang Gong received the B.E. degree in computer science from the University of Science and Technology of China, in 2010, and the Ph.D. degree in computer science from the University of California, Berkeley, in 2015. He is currently an Assistant Professor with the Electrical and Computer Engineering Department and the Computer Science Department (by courtesy) at Iowa State University. He is broadly interested in cybersecurity, privacy, and their intersection with data science. In particular, he has recently been focusing on secure and privacy-preserving social Web services, authentication, and security and privacy in Internet-of-Things.

Prateek Mittal received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, in 2012. He was a Postdoctoral Scholar with the University of California, Berkeley. He is currently an Assistant Professor with the Department of Electrical Engineering, Princeton University. His research interests include the domains of privacy enhancing technologies, trustworthy social systems, and Internet/network security. His work has influenced the design of several widely used anonymity systems, and he is a recipient of several awards, including an ACM CCS outstanding paper. He served as the Program Cochair for the Hot-PETs Workshop in 2013 and 2014.

Raheem Beyah (SM'09) received the B.S. 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 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 the 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. He received the National Science Foundation CAREER Award in 2009, and he was selected for DARPA's Computer Science Study Panel in 2010.