DE-ANONYMIZING SOCIAL NETWORKS AND MOBILITY TRACES

A Thesis
Presented to
The Academic Faculty

by

Weiqing Li

In Partial Fulfillment
of the Requirements for the Degree
Masters of Science in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2016

DE-ANONYMIZING SOCIAL NETWORKS AND MOBILITY TRACES

Approved by:

Dr. Raheem Beyah, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. John Copeland
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Yusun Chang
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 1/15/2016

To

my parents, Chao Li and Bin Wu,

my family, and all my friends,

and last but not the least,

to the Communication Assurance and Performance (CAP) Group,

without whom none of my success would be possible.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Beyah, for his continuous support and guidance. I also thank the members of my thesis committee for the valuable comments and suggestions they have provided. I would like to express my gratitude to all the members of the Communications Assurance and Performance (CAP) group especially Shouling Ji and Shukun Yang without whom this work would not be possible. Lastly, I thank my family and friends for the support they have provided.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

When people utilize social applications and services, their privacy suffers potential serious threats. In this work, we present a novel, robust, and effective de-anonymization attack to mobility trace data and social network data. First, we design a Unified Similarity (US) measurement which takes into account local and global structural characteristics of data, information obtained from auxiliary data, and knowledge inherited from on-going de-anonymization results. By analyzing the measurement on real datasets, we find that some datasets can potentially be de-anonymized accurately and the others can be de-anonymized in a coarse granularity. Utilizing this property, we present a US based De-Anonymization (DA) framework, which iteratively de-anonymizes data with an accuracy guarantee. Then, to de-anonymize large scale data without the knowledge of the overlap size between the anonymized data and the auxiliary data, we generalize DA to an Adaptive De-Anonymization (ADA) framework. By strategically working on two *core matching subgraphs*, ADA achieves high de-anonymization accuracy and reduces computational overhead. Finally, we examine the presented de-anonymization attack on three well known mobility traces: St. Andrews, Infocom06, and Smallblue, and three social network datasets: ArnetMiner, Google+, and Facebook. The experimental results demonstrate that the presented de-anonymization framework is very effective and robust to noise.

CHAPTER I

INTRODUCTION

Social networking services are a fast-growing business nowadays. The development of smart phone technologies further advances the proliferation of social applications and services, such as instant messaging (e.g., IRC, AIM, MSN, Jabber, Skype), sharing sites (e.g., Flickr, Picassa, YouTube, Plaxo), blogs (e.g., Blogger, WordPress, LiveJournal), wikis (e.g., Wikipedia, PBWiki, Wolfram MathWorld), microblogs (e.g., Twitter, Jaiku), social sites (e.g., Facebook, MySpace, Ning, Google+), and collaboration networks (e.g., DBLP, ArnetMiner). Due to the big commercial value to businesses and huge impacts to the society, social networks and data analysis have attracted more and more research interests [2][18][26][11][12][13].

When users participate in online social network activities, e.g., create personal portfolios and connect to friends, or utilize social network functions, e.g., post current location or share information with virtual social friends, people's privacy encounters potential serious threats. On the other hand, to utilize the huge amount of users' data for commercial or academic purposes, social network owners usually release social network data for research (data mining) or transfer data to business partners for target advertising [18]. Furthermore, the advance of mobile computing and communication enables devices such as smartphones to gather user information [26]. For example, users can easily update location, and share posts through Twitter/Facebook on their smartphones.

To protect user privacy, social network owners and services providers usually anonymize data by removing "Personally Identifiable Information (PII)" before releasing the data. However, this anonymized data is still vulnerable to social auxiliary information based data de-anonymization attacks [2][18][26]. The the vulnerability is the results of two fundamental facts. First, when network owners and services providers publish data, only naive anonymization techniques are applied to remove basic PII. For example, the carefully processed and anonymized Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix and was released for the research contest purpose, had enough information in the data's structure that the user's privacy were breached [17]. The second fact is the wide and common availability of social auxiliary information [2][18][26]. As indicated in [18][26], adversaries can obtain social auxiliary information easily or with small efforts through multiple channels, e.g., academic and government data mining, advertising, third-party applications, data aggregation and inferring, privacy attack and acquiring, smart sensing and collection. Even if the availability of large scale auxiliary information is unlikely, a small amount of auxiliary knowledge is usually enough for a successful privacy breach.

A few de-anonymization attacks have been designed for social network data [2][18] or mobility trace data [26]. However, existing works are limited due to one or several reasons, e.g., scalability, generality, and robustness. Our work improves existing works in some or all of the following aspects. First, we significantly improve the de-anonymizaiton accuracy and decrease the computational complexity by proposing a novel *Core Matching Subgraphs* (CMS) based adaptive de-anonymization strategy. Second, besides utilizing node's local property, we incorporate node's global property into de-anonymization without incurring high computational complexity. Furthermore, we also define and apply two new similarity measurements in the proposed de-anonymization technique. Finally, the de-anonymization algorithm presented in this work is a much more general attack framework. It can be applied to both mobility trace data and social network data, directed and undirected data graphs, and weighted and unweighted datasets.

In summary, our main contributions in this thesis are as follows.

1. We analyze three de-anonymization metrics, namely *structural similarity*, *relative distance similarity*, and *inheritance similarity*. By structural similarity, we consider both the local and the global topological characteristics of a node and then quantify the similarity between two nodes with respect to their structural properties. By relative distance similarity, we measure how two nodes are similar from the perspective of auxiliary seed information. By inheritance similarity, we quantify the similarity between two nodes in terms of the knowledge given by nodes who have already been de-anonymized. We also examine how the three measurements function on real datasets. By conducting experiments, we find that some anonymized nodes are significantly distinguishable with respect to some metrics, which suggests that these nodes are potentially easy to de-anonymize. On the other hand, for the other nodes with indistinctive characteristics, they can also be de-anonymized, but with a more coarse granularity.

2. Toward effective de-anonymization, we define a *Unified Similarity* (US) measurement by synthetically considering the defined structural similarity, relative distance similarity, and inheritance similarity. Subsequently, we propose a US based **D**e-**A**nonymization (DA) framework, by which we iteratively de-anonymize the anonymized data with accuracy guarantee provided by a de-anonymization threshold and a mapping control factor.

3. To de-anonymize large-scale data without the knowledge on the overlap size between the anonymized data and the auxiliary data, we generalize DA to an **A**daptive **D**e-**A**nonymization (ADA) framework. ADA adaptively conducts data de-anonymization starting from two *Core Matching Subgraphs* (CMSs), which are defined to estimate the overlap size between the anonymized data and

the auxiliary data. By strategically working on CMSs, the de-anonymization in ADA is limited in two relatively small subgraphs with more information confidence, and thus the de-anonymization accuracy is improved and the computational overhead is reduced. In addition, we also extend DA/ADA to the scenario that the anonymized data or the auxiliary data cannot be modeled by connected graphs.

4. We apply the presented de-anonymization framework to three well known mobility traces: St. Andrews [3], Infocom06 [21], and Smallblue [24]. The experimental results demonstrate that the presented de-anonymization attack is very effective and robust. With only the knowledge of one seed mapping, 57.7%, 93.2%, and 78.3% of the data in St. Andrews, Infocom06, and Smallblue can be successfully de-anonymized, respectively. Furthermore, even when 20% of noise is added into the anonymized data, 80.8%, 50.7%, and 60.8% of the data in St. Andrews, Infocom06, and Smallblue can still be successfully de-anonymized (with five seed mappings).

5. We also examine the presented de-anonymization attack on social network datasets: ArnetMiner (a weighted coauthor dataset consists of 1,127 authors and 6,690 coauthor relationships) Google+ (two datasets with one consists of 5,200 users and 7,062 connections and the other consists of 5,200 users and 7,813 connections), and Facebook (63,731 users and 1,269,502 “friendship” relationships). Again, the experimental results demonstrate the effectiveness and robustness of the presented de-anonymization framework. Based solely on the knowledge of five seed mappings, 96% of users in ArnetMiner (with 4% noise) and 58% of users in Google+ can be successfully de-anonymized. More importantly and surprisingly, even the overlap between the anonymized data and the auxiliary data is just 20% in Facebook, 90.8% of the common users can

also be successfully de-anonymized with false positive error of 8.6% according to 20 seed mappings. Furthermore, we also analyze the impact of *leaf users* (users with one connection) on the de-anonymization performance according to experiments on real data.

The rest of this thesis is organized as follows. In Chapter 2, we survey the most related work. In Chapter 3, we give the preliminaries and considered data model. In Chapter 4, the de-anonymization framework is presented. In Chapter 5, the proposed de-anonymization framework is refined and extended to general large scale social network datasets. We illustrate and discuss the results from extensive experiments on real social and mobility datasets in Chapter 6. Finally, we conclude this thesis in chapter 7.

CHAPTER II

LITERATURE REVIEW

In this chapter, we survey the related work. We first review the specific de-anonymization attacks on and defenses for social and mobility datasets. Then, we discuss the differences that distinguish the proposed de-anonymization attack from existing de-anonymization attacks.

Social and mobility trace data are now easily obtainable and available through multiple channels, e.g., academic and government data mining, advertising, third-party applications, data aggregation and inferring, privacy violating attacks [2][18][26]. To protect the privacy of publicly released data, a common method is to anonymize data by removing PII, e.g., names, age, social security number, before releasing data. However, this naive data anonymization is usually vulnerable to de-anonymization attacks [4][9][15]. Therefore, several further strategies are proposed with the main idea of perturbing the raw data by increasing the automorphism of the data itself, which could make the released data non-distinguishable and thus defend against the de-anonymization attacks. In the following section, we will survey existing anonymization and de-anonymization solutions followed by presenting the merits and differences that distinguish our method from existing works.

2.1 Anonymizing Social and Mobility Data

To preserve the privacy of sensitive relationships in graph data, Zheleva and Getoor designed five different privacy preservation strategies depending on the amount of data removed and the amount of privacy preserved [32]. However, the common availability of auxiliary information for an adversary is not taken into account in the designed strategies.

In [9], Hay et al. introduced k-anonymity to social network data anonymization. An assumption made on adversary's information is that the attacker only has the degree knowledge about the target or partial structural knowledge on the neighborhood of the target. Nevertheless, in reality, the adversary has much more auxiliary information available easily or with a small effort (e.g., through academic and government data mining, advertising, third-party applications). On the other hand, the designed k-anonymity scheme is applicable to low-average-degree social graphs [18]. Nevertheless, the fact is, social graphs' average degree tends to be large and still increasing [16][8]. For instance, the numbers of nodes and edges in a connected component of Google+ are 69,501 and 9,168,660 respectively, which implies a large average degree of 263.8.

In [4], Campan and Truta extended the k-anonymity scheme in [9] by defining an information loss measure that quantifies the amount of structural information loss due to edge generalization. A similar k-anonymity approach is also applied towards ID-anonymization on graphs by Liu and Terzi in [15], where the priori knowledge of adversaries is assumed to be the degree of certain nodes only. As we pointed out before, adversaries can obtain much richer auxiliary information easily or with a small effort. More importantly, as indicated in [18], the cornerstone of k-anonymity is based on data's syntactic property, which may not work on protecting actual data privacy even been satisfied.

2.2 De-anonymizing Social and Mobility Data

The most closely related work to this thesis are [2][18][26][11]. In [2], Backstrom introduced both active attacks and passive attacks to de-anonymize social network 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 (practically and intuitively, it is not straightforward to know when and which part of social data will be released, as well as when to implant Sybil nodes). As analyzed in the subsequent work [18], many reasons limit the practicality of the active attack. A direct limitation is that the active attack is not scalable and difficult to control because the amount of social network data continues to increase [16][8]. To execute an active attack, many Sybil nodes and relationships/ties should be created which is not practical. Furthermore, Sybil defense schemes [1][30][29][31] make this even more difficult. On the other hand, in real online social networks, target nodes have no reason to respond to the connection requests from strange Sybil nodes. For the passive attack in [2], the adversary can breach the privacy of users whom they are linked, which is again suitable for small social networks and difficult to extend to large scale social network data.

In [18], Narayanan and Shmatikov extended the de-anonymization attack to large-scale directed social network data, i.e., the social network data carries direction information which can be used as auxiliary knowledge. The designed de-anonymization algorithm including two phases: *seed identification* and *propagation*. In the seed identification phase, a set of seed mappings are identified between the anonymized graph and the auxiliary graph. In the propagation process, the identified seed mappings are propagated to general mappings between the anonymized graph and the auxiliary graph by employing several heuristic metrics, including eccentricity, edge directionality, node degrees, revisiting nodes, and reverse match. The time complexity of the propagation phase in [18] is $O((|E_1| + |E_2|)d_1 d_2) = O(n^4)$, where $|E_1|$ and d_1 (respectively, $|E_2|$ and d_2) are the edge set cardinality and degree bound of the anonymized graph (respectively, auxiliary graph), respectively, and n is the number of nodes in the anonymized graph or auxiliary graph (same from the order perspective).

In [26], Srivatsa and Hicks presented the first de-anonymization attack to mobility traces while using social networks as a side-channel. The de-anonymization process also consists of two phases: landmark (seed) selection and mapping propagation. In the landmark selection phase, k landmarks with the highest betweenness scores will be selected in the anonymized graph and the auxiliary graph respectively as seeds. In the propagation process, three schemes are developed for graph matching (de-anonymization), namely distance vector, randomized spanning trees, and recursive sub-graph matching. To give a high graph matching (de-anonymization) accuracy, the mapping propagation process will be repeated for each of all the $k!$ possible landmark mappings between the anonymized graph and the auxiliary graph, which is very time-consuming (for example, to de-anonymize the Smallblue dataset which has 125 nodes [24] with 5 landmarks, it takes the designed mapping propagation schemes 6.7 hours, 6.2 hours, and 0.5 hours, respectively). Therefore, scalability could be a significant limitation of the work in [26].

In [11], Ji et al. studied the de-anonymizability of graph data. Specifically, they quantified the structural conditions for perfectly or partially de-anonymizing an anonymized graph. Furthermore, according to the quantification, they proposed a seed-free de-anonymization attack, which is suitable for dense and large-scale graphs. Subsequently, Ji et al. further theoretically studied the de-anonymizability of social networks with seed-knowledge in [12]. They also provided the conditions for perfectly or partially de-anonymizing social networks with seed knowledge. Recently, Ji et al. developed SecGraph, a uniform and open-source platform for graph data anonymization and de-anonymization [13][22]. In SecGraph, they implemented and evaluated 11 graph anonymization schemes, 12 graph utility metrics, 7 application utility metrics, and 15 modern graph de-anonymization attacks (including the two attacks proposed in this thesis). They found that existing anonymization schemes are still vulnerable to one or several de-anonymization attacks. The degree of vulnerability of each anonymization scheme depends on how much and which data utility it preserves. In [20], Nilizadeh et al. studied how to use the graph's community information to enhance existing seed-based de-anonymization attacks, e.g., [18][26]. They proposed a community-based de-anonymization framework, which de-anonymizes a graph first at the community-level and then at the user-level.

2.3 Remark

Some or all of the following aspects distinguish this work from existing techniques. First, when de-anonymizing large datasets, we define CMS in both the anonymized graph and the auxiliary graph according to seed information. Based on CMS, we propose a novel adaptive de-anonymization strategy which is quite suitable for large scale data de-anonymization. Following this strategy, we de-anonymize the nodes in CMS first and then propagate the de-anonymization by spanning CMS in both graphs adaptively. In this manner, we can significantly improve the de-anonymization accuracy and decrease the computational complexity. Second, the degree centrality can only indicate the local property of a node in a graph. In some anonymized data, the fact that many nodes with similar degrees blurs or even invalidates the effectiveness of using degree to match/distinguish nodes. Therefore, we include metrics indicating global properties of a node in a graph into our consideration, e.g., closeness centrality, betweenness centrality. Furthermore, besides utilizing structural knowledge, we also define and apply two similarity measurements in the proposed de-anonymization technique, namely the *relative distance similarity* and the *inheritance similarity.* This increases the de-anonymization efficiency and accuracy. More importantly, the computational cost induced by including new global metrics can be overcome through the CMS-based adaptive de-anonymization in large scale datasets. Third, the de-anonymization attack presented in [18] applies to social network data that can be modeled by directed graphs, where the direction information is assumed to be free auxiliary information for adversaries. In this work, we consider a more general scenario by removing the direction limitation. Our de-anonymization algorithm works for undirected graphs as well as directed graphs by incorporating the direction heuristic in [18]. Finally, we also consider the potential weight (relationship strength) information on edges of anonymized graphs. Therefore, our de-anonymization algorithm is also effective on weighted graphs. In summary, the de-anonymization attack presented in this work applies to large scale social network data, mobility trace data, directed/undirected data graphs and weighted/unweighted data graphs, and is more general than previous works.

CHAPTER III

PRELIMINARIES AND SYSTEM MODEL

3.1 Anonymized Data Graph

In this thesis, we consider the anonymized data which can be modeled by an undirected graph[1], denoted by $G^a = (V^a, E^a, W^a)$, where $V^a = \{i | i \text{ is a node}\}$ is the node set (e.g., users in an anonymized Google+ graph), $E^a = \{l^a_{i,j} | i, j \in V^a, \text{ and there is tie between } i \text{ and } j\}$ is the set of all the links existing between any two nodes in V^a (a link could be a friend relationship such as in Google+ or a contact relationship such as in the mobility trace St Andrew), and $W^a = \{w^a_{i,j} | i, j \in V^a, l^a_{i,j} \in E^a, w^a_{i,j} \text{ is a real number}\}$ is the set of possible weights associated with links in E^a (e.g., in a coauthor graph, the weight of a coauthor relationship could be the number of coauthored papers). If G^a is an unweighted graph, we simply define $w^a_{i,j} = 1$ for each link $l^a_{i,j} \in E^a$.

For $\forall i \in V^a$, we define its neighbor set as $N^a(i) = \{j \in V^a | l^a_{ij} \in E^a\}$. Then, $\Delta^a_i = |N^a(i)|$ represents the number of neighbors of i in G^a. For $\forall i, j \in V^a$, let $p^a(i, j)$ be a shortest path from i to j in G^a and $|p^a(i, j)|$ be the number of links on $p^a(i, j)$ (the number of links passed from i to j through $p^a(i, j)$). Then, we define $\mathbb{P}^a_{i,j} = \{p^a(i, j)\}$ the set of all the shortest pathes between i and j. Furthermore, we define the diameter of G^a as $D^a = \max\{|p^a(i, j)| \forall i, j \in V^a, p^a(i, j) \in \mathbb{P}^a_{i,j}\}$, i.e. the length of the longest shortest path in G^a.

[1] Note that, the de-anonymization algorithm designed in this thesis can also be applied to directed graphs directly by overlooking the direction information on edges, or by incorporating the edge-direction based de-anonymizatoin heuristic in [18].

3.2 Auxiliary Data Graph

As in [18][26], we assume the auxiliary data is the information crawled in current online social networks, e.g., the "follow" relationships on Twitter [18], the "contact" relationships on Flickr, the "friend" relationships on Facebook, the "circle" relationships on Google+. Furthermore, similar as the anonymized data, the auxiliary data can also be modeled as an undirected graph $G^u = (V^u, E^u, W^u)$, where V^u is the node set, E^u is set of all the links (relationships) among the nodes in V^u, and W^u is the set of possible weights associated with the links in E^u. As the definitions on the anonymized graph G^a, we can define the neighborhood of $\forall i \in V^u$ as $N^u(i)$, the shortest path set between $i \in V^u$ and $j \in V^u$ as $\mathbb{P}^u(i,j) = \{p^u(i,j)\}$, and the diameter of G^u as $D^u = \max\{|p^u(i,j)| \forall i, j \in V^u, p^u(i,j) \in \mathbb{P}^u(i,j)\}$.

In addition, we assume G^a and G^u are connected. Note that this is not a limitation of our scheme. The designed de-anonymization algorithm is also applicable to the case where G^a and G^u are not connected. We will discuss this in Chapter 5.

3.3 Attack Model

Our de-anonymization objective is to map the nodes in the anonymized graph G^a to the nodes in the auxiliary graph G^u as accurate as possible. Then, adversaries can rely on the auxiliary data such as the portfolio created by users in online social networks to breach users' privacy. Formally, let $\gamma(v)$ be the *objective reality* of $v \in G^a$ in the physical world. Then, an ideal de-anonymization can be represented by mapping $\Phi : G^a \to G^u$, such that for $v \in G^a$,

$$\Phi(v) = \begin{cases} v', & \text{if } v' = \Phi(v) \in V^u; \\ \perp, & \text{if } \Phi(v) \notin V^u. \end{cases} \tag{1}$$

where $\perp$ is a special *not existing indicator.* Now, let

$$\mathcal{M} = \{(v_1, v_1'), (v_2, v_2'), \cdots, (v_n, v_n')\}$$

be the outcome of a de-anonymization attack such that

$$\begin{cases} v_i \in V^a, \cup v_i = V^a, n = |V^a|, & i = 1, 2, \cdots, n; \\ v'_i = \Phi(v_i), v'_i \in V^u \cup \{\perp\}, & i = 1, 2, \cdots, n. \end{cases} \tag{2}$$

Then, the de-anonymization on v_i is said to be *successful* if

$$\begin{cases} \Phi(v_i) = \gamma(v_i), & \text{if } \gamma(v_i) \in V^u; \\ \Phi(v_i) = \perp, & \text{if } \gamma(v_i) \notin V^u. \end{cases} \tag{3}$$

or *failure* if

$$\begin{cases} \Phi(v_i) \in \{u | u \in V^u, u \neq \gamma(v_i)\} \cup \{\perp\}, & \text{if } \gamma(v_i) \in V^u; \\ \Phi(v_i) \neq \perp, & \text{if } \gamma(v_i) \notin V^u. \end{cases} \tag{4}$$

In this thesis, we are aiming to design a de-anonymization framework with a high success rate (accuracy) while a low failure rate. In addition, the designed de-anonymization algorithm is expected to be robust to noise and scalable to large scale datasets.

3.4 Datasets

In this thesis, we employ six well known datasets to examine the effectiveness of the designed de-anonymization framework: St. Andrews/Facebook [3][26], Infocom06/DBLP [21][26], Smallbule/Facebook [24][26], ArnetMiner [27], Google+ [8], and Facebook [28]. St. Andrews, Infocom06, and Smallbule are three mobility trace datasets. The St. Andrews dataset contains WiFi-recorded mobility trace data of 27 T-mote users through 30 days deployed in the University of St. Andrews. The Infocom06 trace includes Bluetooth sightings by a group of 78 users carrying iMotes for four days in IEEE INFOCOM 2005 in the Grand Hyatt Miami. The Smallbule dataset consists of contacts among 125 instant messenger users on an enterprise network. An overview of the three mobility traces is shown in Table 1. We employ the exact same techniques as in the previous work [26] to preprocess the three mobility trace datasets to obtain three anonymized data graphs. To de-anonymize the aforementioned three

Table 1: Overview of St. Andrews, Infocom06, Smallblue, and associated social networks.

	St. Andrews	**Infocom06**	**Smallblue**
Comm. network type	WiFi	Bluetooth	IM
Comm. nodes No.	27	78	125
Duration (days)	30	4	30
Granularity (secs)	300	120	300
Contacts No.	18,241	182,951	240,665
Social network type	Facebook	DBLP	Facebook
Social nodes No.	27	616	400

anonymized mobility data traces, we employ three auxiliary social network datasets [26] associated with these three mobility traces. For the St. Andrews dataset, we have a Facebook dataset indicating the "friend" relationships among the T-mote users in the trace. For the Infocom06 dataset, we employ a coauthor dataset consisting of 616 authors obtained from DBLP which indicates the "coauthor" relationships among all the attendees of INFOCOM 2005. For the Smallblue dataset, we have a Facebook dataset indicating the "friend" relationships among 400 employees from the same enterprise as Smallblue. Note that, the social network datasets corresponding to Infocom06 and Smallblue are supersets of them.

We also apply the presented de-anonymization attack to social network datasets: ArnetMiner [27], Google+ [7], and Facebook [28]. ArnetMiner is an online academic social network. In this thesis, the employed data is extracted from ArnetMiner in 2011 on topic "Database Systems / XML Data" which consists of 1,127 authors and 6,690 "coauthor" relationships. For each coauthor relationship, there is a weight associated with it indicating the number of coauthored papers by the two authors. Consequently, the ArnetMiner data can be modeled by a weighted graph. Furthermore, we know the ground truth of the ArnetMiner data. When using it to examine the presented de-anonymization attack, we will anonymize it first by adding different levels of noise. Then, we apply our method to de-anonymize it. As a new social network, Google+ was launched in early July 2011. We use two Google+ datasets which were created on July 19 and August 6 in 2011 [7], denoted by JUL and AUG respectively. Both JUL and AUG consist of 5,200 users as well as their profiles. In addition, there were 7,062 connections in JUL and 7,813 connections in AUG. By insight analysis [7], some connections appeared in AUG may not appear in JUL and vise versa. This is because a user may add new connections or disable existing connections. Furthermore, the two datasets are preprocessed as undirected graphs. Since we know the hand labeled ground truth of JUL and AUG, we will examine the presented de-anonymization framework by de-anonymizing JUL with AUG as auxiliary data and then de-anonymizing AUG with JUL as auxiliary data. The Facebook dataset consists of 63,731 users and 1,269,502 “friend” relationships (links). To use this dataset to examine the presented de-anonymization attack, we will preprocess it based on the known hand labeled ground truth. For more detailed experimental settings and data processing, we will describe them in the experimental chapter (Chapter 6).

CHAPTER IV

DE-ANONYMIZATION

From a macroscopic view, the designed de-anonymization attack framework consists of two phases: seed selection and mapping propagation. In the seed selection phase, we identify a small number of seed mappings from the anonymized graph G^a to the auxiliary graph G^u serving as landmarks to bootstrap the de-anonymization. In the mapping propagation phase, we de-anonymize G^a through synthetically exploiting multiple similarity measurements. Since seed selection can be implemented by many existing strategies and will not be our primary technical contribution, we will discuss it briefly and focus on how to design an effective mapping propagation scheme.

4.1 Seed Selection and Mapping Spanning

The rational and feasibility of seed selection in our de-anonymization framework (as well as other de-anonymization attacks) lies in three realities. The first is the common availability of huge amounts of social network data, which is an open and rich source for obtaining a small number of seeds. For instance, (*i*) for the data published for academic and government data mining, some auxiliary information may be released at the same time or can be obtained easily [17]; (*ii*) the social network data (e.g., Facebook, MySpace, Google) shared with advertising partners by social network operators may cause some information leakage, which could be used as auxiliary seed data for de-anonymization attacks [18]; (*iii*) online social network operators (e.g., Facebook, Twitter) and researchers (e.g., Stanford SNAP Datasets [25], Dartmouth CRAWDAD [5]) publish many kinds of anonymized/unanoymized social network data periodically; etc. The second reality is the existence of multiple effective channels to obtain a small number of seed mappings (actually, we can obtain much richer auxiliary information). Some example channels are as follows: (*i*) seed mapping information could be acquired due to data leakage, e.g., some data may be leaked in data release for academic and government data mining with the original purpose of assisting research [18][17]; (*ii*) auxiliary information can be collected by launching third-party applications on online social networks (many successful examples are surveyed in [18]); (*iii*) considering the common availability of huge amounts of social network data, another effective method to infer auxiliary information is by data aggregation. Especially in current online social networks, the degree distribution of nodes (corresponding to users) has been shown to follow the power law distribution in many cases. Therefore, important nodes could be inferred easily and accurately in terms of their centrality [26]; (*iv*) it is also possible to obtain a small number of seed mappings in a human-assisted semi-automatic manner. Adversaries can crawl some data first and then relies on human-assisted semi-automatic analysis to obtain some auxiliary information [17]; etc. The third reality is that a small number of seed mappings is sufficiently helpful (or enough depends on the required accuracy) to our de-anonymization framework. As shown in our experiments, a small number of seed mappings (sometimes even one seed mapping) are sufficient to achieve highly accurate de-anoymization.

In our de-anonymization framework, we can select a small number of seed mappings by employing multiple seed selection strategies [2][18][26] individually or collaboratively. Some candidate seed selection strategies are as follows. (1) One method to obtain a small number of seed mappings can be implemented by a Sybil attack [2], in which some Sybil nodes will be implanted into the target social network. Then, we can use the social neighbors of the Sybil nodes or the Sybil nodes themselves as seeds. Although large-scale Sybil attack to a network is difficult [29][31], local or small scale Sybil attack to obtain some seed mappings is practical. (2) Another applicable method to obtain a small number of seed mappings is by compromising nodes [2][18][26]. An adversary could collude with some users in the anonymized data to obtain some seed mapping information. In addition, the adversary himself could be some node in the anonymized graph. In this case, it is even easier to obtain seed mapping information. (3) As we analyzed before, seed mappings can also be obtained by launching third-party applications on the target network (e.g., Facebook, Twitter). Again, it may be impossible to collect auxiliary information in large scale, however, small scale of auxiliary information collection for seed mapping is practical [23][10][6]. (4) Some other existing attacks and seed identifying algorithms can be employed for seed selection, e.g., the seed selection used in [2] for active and passive attacks, the clique based seed identification in [18].

Since seed selection is not the primary contribution in this thesis, we assume we have identified κ seed mappings by exploiting the aforementioned strategies individually or collaboratively, denoted by $\mathcal{M}_s = \{(s_1, s'_1), (s_2, s'_2), \cdots, (s_\kappa, s'_\kappa)\}$, where $s_i \in V^a$, $s'_i \in V^u$, and $s'_i = \Phi(s_i)$. In the mapping propagation phase, we will start with the seed mapping $\mathcal{M}_s$ and propagate the mapping (de-anonymization) to the entire G^a iteratively. Let $\mathcal{M}_0 = \mathcal{M}_s$ be the initial mapping set and $\mathcal{M}_k$ $(k = 1, 2, \cdots)$ be the mapping set after the k-th iteration. To facilitate our discussion, we first define some terminologies as follows.

Let $M_k^a = \bigcup\limits_{i=1}^{|\mathcal{M}_k|} \{v_i | (v_i, v'_i) \in \mathcal{M}_k\}$ and $M_k^u = \bigcup\limits_{i=1}^{|\mathcal{M}_k|} \{v'_i | (v_i, v'_i) \in \mathcal{M}_k\} \setminus \{\perp\}$ be the sets of nodes that have been mapped till iteration k in G^a and G^u, respectively. Then, we define the 1-hop mapping spanning set of M_k^a as $\Lambda^1(M_k^a) = \{v_j \in V^a | v_j \notin M_k^a$ and $\exists v_i \in M_k^a$ $s.t.$ $v_j \in N^a(v_i)\}$, i.e., $\Lambda^1(M_k^a)$ denotes the set of nodes in G^a that have some neighbor been mapped and themselves not been mapped yet. To be general, we can also define the δ-hop mapping spanning set of M_k^a as $\Lambda^\delta(M_k^a) = \{v_j \in V^a | v_j \notin M_k^a$ and $\exists v_i \in M_k^a$ $s.t.$ $|p^a(v_i, v_j)| \leq \delta\}$, i.e., $\Lambda^\delta(M_k^a)$ denotes the set of nodes in G^a that are at most δ hops away from some node been mapped and themselves not been mapped yet. Here, $\delta (\delta = 1, 2, \cdots)$ is called the spanning factor in the mapping

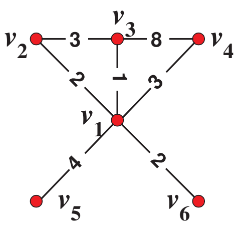


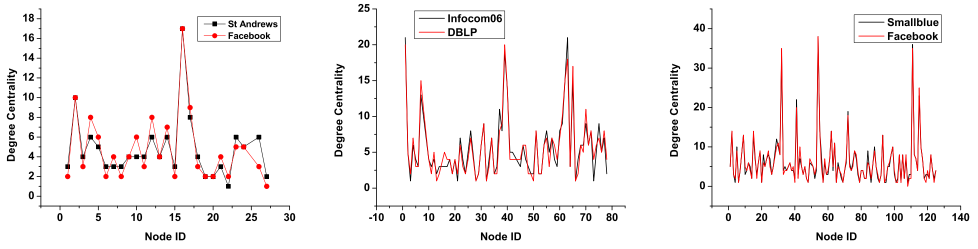
Figure 1: A weighted graph.

propagation phase of the proposed de-anonymization framework. Similarly, we can define the 1-hop mapping spanning set and δ-hop mapping spanning set for M_k^u as $\Lambda^1(M_k^u) = \{v'_j \in V^u | v'_j \notin M_k^u$ and $\exists v'_i \in M_k^u$ $s.t.$ $v'_j \in N^u(v'_i)\}$ and $\Lambda^\delta(M_k^u) = \{v'_j \in V^u | v'_j \notin M_k^u$ and $\exists v'_i \in M_k^u$ $s.t.$ $|p^u(v'_i, v'_j)| \leq \delta\}$, respectively.

Based on the defined δ-hop mapping sets $\Lambda^\delta(M_k^a)$ and $\Lambda^\delta(M_k^u)$, we try to seek a mapping Φ which maps the anonymized nodes in $\Lambda^\delta(M_k^a)$ to some nodes in $\Lambda^\delta(M_k^u) \cup \{\perp\}$ iteratively in the mapping propagation phase of our de-anonymization framework. To make the mapping propagation phase effective and controllable, we define several important measurements according to nodes' local properties, global properties, relative global property as well as inheritance properties in the following sections before giving the de-anonymization framework.

4.2 Structural Similarity

In graph theory, the concept of centrality is often used to measure the topological importance and characteristic of a node within a graph. In this thesis, we employ three centrality measurements to capture the topological property of a node in G^a or G^u, namely degree centrality, closeness centrality, and betweenness centrality. In the case that the considering data is modeled by a weighted graph, we also defined the weighted version of the employed three centrality measurements.



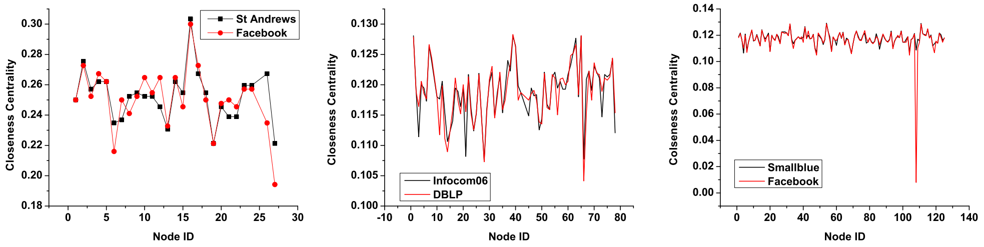
(a) St. Andrews vs Facebook (b) Infocom06 vs DBLP (c) Smallblue vs Facebook

Figure 2: Degree centrality.

4.2.1 Degree Centrality and Weighted Degree Centrality

The degree centrality is defined as the number of ties that a node has in a graph, i.e., the number of links with this node as an endpoint. For instance, in the considering anonymized data graph, the degree centrality of $v \in V^a$ is defined as $d_v = \Delta_v^a = |N^a(v)|$. Similarly, for $v' \in V^u$, its degree centrality is $d_{v'} = \Delta_{v'}^u = |N^u(v')|$. We calculate the degree centrality of the nodes in St. Andrews, Infocom06, and Smallblue, as well as their counterparts in the corresponding social graphs (Facebook, DBLP, and Facebook), and the results are shown in Fig.2. From Fig.2, we observe that the degree centrality distributions of the anonymized graph and auxiliary graph are similar, which implies that degree centrality can be used for de-anonymization. On the other hand, multiple nodes in both graphs may have similar degree centrality, which suggests that degree centrality as a structural measurement can be used for coarse granularity de-anonymization.

When the data being considered is modeled by a weighted graph as shown in Fig.1, which consists of 6 nodes and 7 links, the weights on links provide extra information in characterizing the centrality of a node. In this case, the degree centrality defined for unweighted graphs cannot properly reflect a nodes' structural importance [19]. For instance, $d_{v_2} = d_{v_4}$ in Fig.1. However, the links associated with v_2 and v_4 have different weights or sum weights, which cannot be reflected by d_{v_2} and d_{v_4}. One naive idea is to define the degree centrality of a node in a weighted graph as the sum of



(a) St. Andrews vs Facebook (b) Infocom06 vs DBLP (c) Smallblue vs Facebook

Figure 3: Closeness centrality.

the weights on the links associated with that node [19]. Nevertheless, this definition overlooks the information about the number of links associated with a node on the other hand. As shown in Fig.1, $\sum\limits_{j\neq 1} w_{1,j} = \sum\limits_{j\neq 3} w_{3,j} = 12$ while $d_{v_1} \neq d_{v_3}$ (as defined in Chapter 3, $w_{i,j}$ is the weight on the link from i to j or 0 if there is no link). To consider both the number of links associated with a node and the weights on these links, we define the *weighted degree centrality* for $v \in V^a$ as

$$wd_v = \Delta_v^a \cdot \left(\frac{\sum\limits_{u \in N^a(v)} w_{v,u}^a}{\Delta_v^a}\right)^{\alpha}, \tag{5}$$

where α is a positive tuning parameter that can be set according to the research setting and data [19]. Basically, when $0 \leq \alpha \leq 1$, high degree is considered more important, whereas when $\alpha \geq 1$, weight is considered more important. Similarly, we can define the *weighted degree centrality* for $v' \in V^u$ as

$$wd_{v'} = \Delta_{v'}^u \cdot \left(\frac{\sum\limits_{u' \in N^u(v')} w_{v',u'}^u}{\Delta_{v'}^u}\right)^{\alpha}. \tag{6}$$

4.2.2 Closeness Centrality and Weighted Closeness Centrality

From the definition of degree centrality, it indicates the local property of a node since only the adjacent links are considered. To fully characterize a node's topological importance, some centrality measurements defined from a global view are also important and useful. One manner to count a node's global structural importance is by *closeness centrality*, which measures how close a node is to other nodes in a graph and is defined as the ratio between $n-1$ and the sum of its distances to all other nodes. In the definition, n is the number of nodes and *distance* is the length in terms of hops from a node to another node in a graph. Formally, for $v \in V^a$, its *closeness centrality* c_v is defined as

$$c_v = \frac{|V^a| - 1}{\sum\limits_{u \in V^a, u \neq v} |p^a(v,u)|}. \tag{7}$$

Similarly, the *closeness centrality* $c_{v'}$ of $v' \in V^u$ is defined as

$$c_{v'} = \frac{|V^u| - 1}{\sum\limits_{u' \in V^u, u' \neq v'} |p^u(v',u')|}. \tag{8}$$

Fig.3 demonstrates the closeness centrality score of the nodes in St. Andrews, Infocom06, and Smallblue, as well as their counterparts in the corresponding social graphs (Facebook, DBLP, and Facebook), respectively. From Fig.3, the closeness centrality distribution of nodes in the anonymized graph generally agrees with that in the auxiliary graph, which suggests that closeness centrality can be a measurement for de-anonymization. In the case that the data being considered is modeled by a weighted graph, we define the *weighted closeness centrality* for $v \in V^a$ and $v' \in V^u$ as

$$wc_v = \frac{|V^a| - 1}{\sum\limits_{u \in V^a, u \neq v} |p^a_w(v,u)|} \tag{9}$$

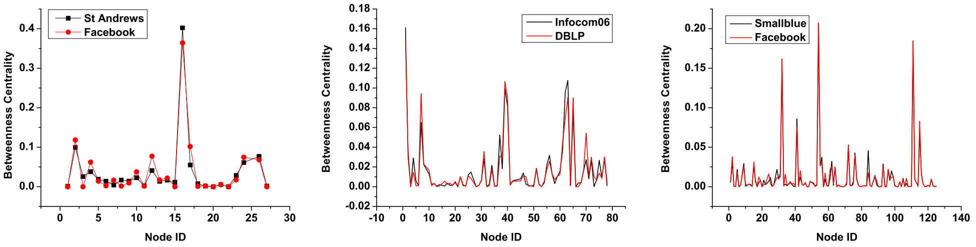
and

$$wc_{v'} = \frac{|V^u| - 1}{\sum\limits_{u' \in V^u, u' \neq v'} |p^u_w(v',u')|}, \tag{10}$$

respectively, where $p^a_w(\cdot,\cdot)/p^u_w(\cdot,\cdot)$ is the shortest path between two nodes in a weighted graph.

4.2.3 Betweenness Centrality and Weighted Betweenness Centrality

Besides closeness centrality, betweenness centrality is another measure indicating a node's global structural importance within a graph, which quantifies the number of



(a) St. Andrews vs Facebook (b) Infocom06 vs DBLP (c) Smallblue vs Facebook

Figure 4: Betweenness centrality.

times a node acts as a bridge (intermediate node) along the shortest path between two other nodes. Formally, for $v \in V^a$, its *betweenness centrality* b_v in G^a is defined as

$$b_v = \frac{\sum\limits_{x \neq v \neq y} \frac{\sigma^a_{xy}(v)}{\sigma^a_{xy}}}{\binom{|V^a|-1}{2}} = \frac{2}{(|V^a|-1)(|V^a|-2)} \cdot \sum_{x \neq v \neq y} \frac{\sigma^a_{xy}(v)}{\sigma^a_{xy}}, \tag{11}$$

where $x', y' \in V^a$, $\sigma^a_{xy} = |\mathbb{P}^a(x, y)|$ is the number of all the shortest paths between x and y in G^a, and $\sigma^a_{xy}(v) = |\{p^a(x, y) \in \mathbb{P}^a(x, y)|v$ is an intermediate node on path $p^a(x, y)\}|$ is the number of shortest paths between x and y in G^a that v lies on. Similarly, the betweenness centrality $b_{v'}$ of $v' \in V^u$ in G^u is defined as

$$b_{v'} = \frac{\sum\limits_{x' \neq v' \neq y'} \frac{\sigma^u_{x'y'}(v')}{\sigma^u_{x'y'}}}{\binom{|V^u|-1}{2}} = \frac{2}{(|V^u|-1)(|V^u|-2)} \cdot \sum_{x' \neq v' \neq y'} \frac{\sigma^u_{x'y'}(v')}{\sigma^u_{x'y'}}. \tag{12}$$

According to the definition, we obtain the betweenness centrality of nodes in St. Andrews/Facebook, Infocom06/DBLP, and Smallblue/Facebook as shown in Fig.4. From Fig.4, the nodes in G^a and their counterparts in G^u agree highly on betweenness centrality. Consequently, betweenness centrality can also be employed in our de-anonymization framework for distinguishing mappings. For the case that the considering data is modeled as a weighted graph, we define the *weighted betweenness centrality* for $v \in V^a$ and $v' \in V^u$ as

$$wb_v = \frac{\sum\limits_{x \neq v \neq y} \frac{\sigma^{wa}_{xy}(v)}{\sigma^{wa}_{xy}}}{\binom{|V^a|-1}{2}} = \frac{2}{(|V^a|-1)(|V^a|-2)} \cdot \sum_{x \neq v \neq y} \frac{\sigma^{wa}_{xy}(v)}{\sigma^{wa}_{xy}} \tag{13}$$

and

$$wb_{v'} = \frac{\sum_{x'\neq v'\neq y'} \frac{\sigma^{wu}_{x'y'}(v')}{\sigma^{wu}_{x'y'}}}{\binom{|V^u|-1}{2}} = \frac{2}{(|V^u|-1)(|V^u|-2)} \cdot \sum_{x'\neq v'\neq y'} \frac{\sigma^{wu}_{x'y'}(v')}{\sigma^{wu}_{x'y'}}, \tag{14}$$

respectively, where σ^{wa}_{xy} and $\sigma^{wa(v)}_{xy}$ (respectively, $\sigma^{wu}_{x'y'}$ and $\sigma^{wa(v')}_{x'y'}$) are the number of shortest paths between x and y (respectively, x' and y') and the number of shortest paths between x and y (respectively, x' and y') passing v (respectively, v') in the weighted graph G^a (respectively, G^u), respectively.

4.2.4 Structural Similarity

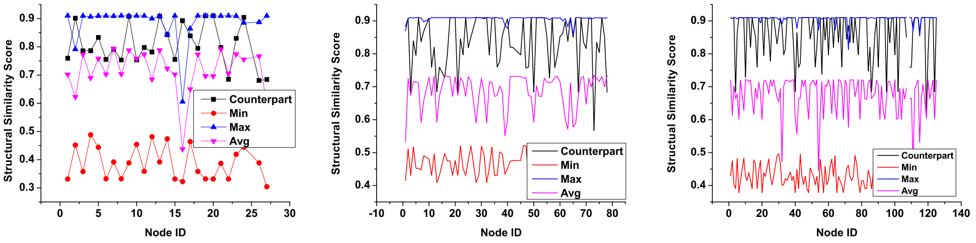
From the analysis on real datasets, the local and global structural characteristics carried by degree, closeness, and betweenness centralities of nodes can guide our de-anonymization framework design. Following this direction, to consider and utilize nodes' structural property integrally, we define a unified structural measurement, namely structural similarity, to jointly count two nodes' both local and global topological properties. First, for $v \in V^a$ and $v' \in V^u$, we define two structural characteristic vectors $\mathbf{S}^a(v)$ and $\mathbf{S}^u(v')$ respectively in terms of their (weighted) degree, closeness, and betweenness centralities as follows:

$$\mathbf{S}^a(v) = [d_v, c_v, b_v, wd_v, wc_v, wb_v] \tag{15}$$

$$\mathbf{S}^u(v') = [d_{v'}, c_{v'}, b_{v'}, wd_{v'}, wc_{v'}, wb_{v'}]. \tag{16}$$

In $\mathbf{S}^a(v)$, if G^a is unweighted, we set $wd_v = wc_v = wb_v = 0$; otherwise, we first count d_v, c_v, and b_v by assuming G^a is unweighted, and then count wd_v, wc_v, and wb_v in the weighted G^a. We also apply the same method to obtain $\mathbf{S}^u(v')$ in G^u. Based on $\mathbf{S}^a(v)$ and $\mathbf{S}^u(v')$, we define the structural similarity between $v \in V^a$ and $v' \in V^u$, denoted by $s_S(v, v')$, as the cosine similarity between $\mathbf{S}^a(v)$ and $\mathbf{S}^u(v')$, i.e.,

$$s_S(v, v') = \frac{\mathbf{S}^a(v) \cdot \mathbf{S}^u(v')}{\|\mathbf{S}^a(v)\|\|\mathbf{S}^u(v')\|}, \tag{17}$$



(a) St. Andrews vs Facebook (b) Infocom06 vs DBLP (c) Smallblue vs Facebook

Figure 5: Structural similarity.

where $\cdot$ is the dot product and $\|\cdot\|$ is the magnitude of a vector.

The structural similarity between the nodes in St. Andrews, Infocom06, and Smallblue and their corresponding auxiliary networks is shown in Fig.5, where Counterpart represents $s_S(v, v' = \gamma(v))$ indicating the structural similarity between $v \in V^a$ and its objective reality $\gamma(v)$ in G^u, *Min* represents $\min\{s_S(v, x')|x' \in V^u, x' \neq \gamma(v)\}$, *Max* represents $\max\{s_S(v, x')|x' \in V^u, x' \neq \gamma(v)\}$, and *Avg* represents $\frac{1}{|V^u|-1} \sum\limits_{x' \in V^u, x' \neq \gamma(v)} s_S(v, x')$. From Fig.5, we have the following two basic observations.

- For some nodes with distinguished structural characteristics, e.g., nodes 2, 16, 24 in St. Andrews, nodes 10, 40 in Infocom06, and nodes 19, 54, 64, 72, 111, 115 in Smallblue, they agree with their counterparts and disagree with other nodes in the auxiliary graphs significantly (actually, they also show the similar agreeableness and disagreeableness with respect to degree, closeness, and betweenness centralities). Consequently, this suggests that these nodes can be de-anonymized even just based on their structural characteristics. In addition, this confirms that structural properties can be employed in de-anonymization attacks.

- For the nodes with indistinctive structural similarities, e.g., nodes 7, 10, 22, 26 in St. Andrews, nodes 16, 73, 78 in Infocom06, and nodes 4, 40, 86, 102, 124 in Smallblue, exact node mapping relying on structural property alone is difficult or impossible to achieve from the view of graph theory. Fortunately, even if this is true, structural characteristics can also help us to differentiate these indistinctive nodes from most of the other nodes in the auxiliary graph. Hence, structural similarity based coarse granularity de-anonymization is practical.

4.3 Relative Distance Similarity

In Section 4.1, we select an initial seed mapping $\mathcal{M}_0 = \mathcal{M}_s = \{(s_1, s'_1), (s_2, s'_2), \cdots, (s_\kappa, s'_\kappa)\}$. This apriori knowledge can be used to conduct more confident ratiocination in de-anonymization. Therefore, for $v \in V^a \setminus M_0^a$, we define its *relative distance vector*[1], denoted by $\mathbf{D}^a(v)$ to the seeds in $M_0^a = \{s_1, s_2, \cdots, s_\kappa\}$ as

$$\mathbf{D}^a(v) = [D_1^a(v), D_2^a(v), \cdots, D_\kappa^a(v)], \tag{18}$$

where $D_i^a(v) = \frac{|p^a(v,s_i)|}{D^a}$ is the normalized relative distance between v and seed s_i. Similarly, based on the initial seed set $M_0^u = \{s'_1, s'_2, \cdots, s'_\kappa\}$ in G^u, we can define the relative distance vector for $v' \in V^u \setminus M_0^u$ to the seeds in M_0^u as

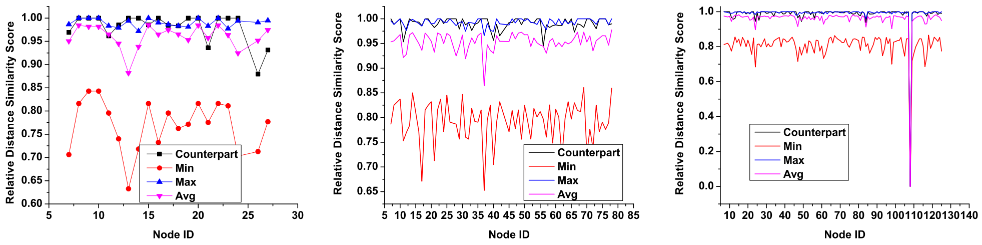
$$\mathbf{D}^u(v') = [D_1^u(v'), D_2^u(v'), \cdots, D_\kappa^u(v')], \tag{19}$$

where $D_i^u(v') = \frac{|p^u(v',s'_i)|}{D^u}$ is the normalized relative distance between v' and seed s'_i. Again, we can define the relative distance similarity between $v \in V^a \setminus M_0^a$ and $v' \in V^u \setminus M_0^u$, denoted by $s_D(v, v')$, as the cosine similarity between $\mathbf{D}^a(v)$ and $\mathbf{D}^u(v')$, i.e.,

$$s_D(v, v') = \frac{\mathbf{D}^a(v) \cdot \mathbf{D}^u(v')}{\|\mathbf{D}^a(v)\| \|\mathbf{D}^u(v')\|}. \tag{20}$$

For St. Andrews/Facebook, Infocom06/DBLP, and Smallblue/Facebook, by assuming $\mathcal{M}_s = \{(i,i)|i = 1, 2, \cdots, 6\}$ (which implies $M_0^a = M_0^u = \{1, 2, 3, 4, 5, 6\}$), we can obtain the relative distance similarity scores between the nodes in $V^a \setminus M_0^a$ and

[1] Note that, the relative distance vector can also be defined using the Multidimensional Scaling (MDS) theory [14]. To consistent with existing anonymization/de-anonymization literature [9][11], we still use the "relative distance (similarity)" term. Mathematically, the used relative distance similarity can be considered as a special case/application of MDS.



(a) St. Andrews vs Facebook (b) Infocom06 vs DBLP (c) Smallblue vs Facebook

Figure 6: Relative distance similarity.

the nodes in $V^u \setminus M_0^u$ as shown in Fig.6. From Fig.6, we can observe the following facts.

- Some anonymized nodes (which may be indistinctive with respect to structural similarity), e.g., nodes 14, 19, 23 in St. Andrews, nodes 28, 31, 37, 54 in Infocom06, and nodes 46, 63, 75, 98, 105 in Smallblue, highly agree with their counterparts and meanwhile disagree with other nodes in the auxiliary graph, which suggests that they can be de-anonymized successfully with a high probability by employing the relative distance similarity based metric.

- For some nodes, e.g., nodes 11, 21, 26, 27 in St. Andrews, nodes, 56, 69 in Infocom06, and nodes 12, 13, 26 in Smallblue, they are indistinctive on the relative distance similarity with respect to the initial seed selection $\{1, 2, 3, 4, 5, 6\}$. To distinguish them, extra effort is expected, e.g., by utilizing structural similarity collaboratively, employing another seed selection.

- The nodes that are significantly distinguishable with respect to structural similarity may be indistinctive with respect to relative distance similarity, and vice versa. This inspires us to design a proper and effective multi-measurement based de-anonymization framework.

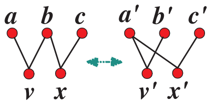


Figure 7: Mapping inheritance.

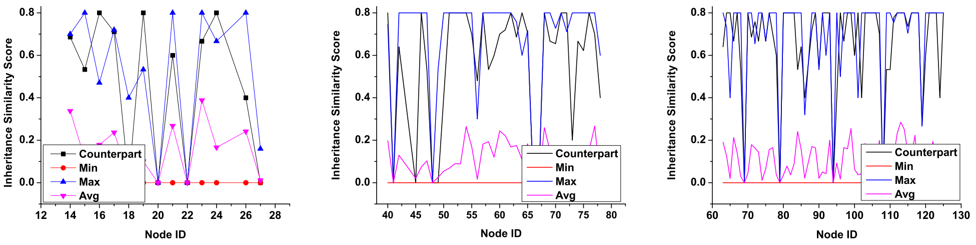
4.4 Inheritance Similarity

Besides the initial seed mapping, the de-anonymized nodes during each iteration, i.e., $\mathcal{M}_k$, could provide further knowledge when de-anonymizing $\Lambda^\delta(M_k^a)$. As shown in Fig.7, if the current de-anonymization result is $\mathcal{M}_k = \{(a, a'), (b, b'), (c, c')\}$, then $\mathcal{M}_k$ can serve as a reference in the next iteration of de-anonymization, i.e., $\mathcal{M}_k$ can provide knowledge to de-anonymize $\Lambda^1(M_k^a) = \{v, x\}$ (assume $\delta = 1$). Therefore, for $v \in \Lambda^\delta(M_k^a)$ and $v' \in \Lambda^\delta(M_k^u)$, we define the knowledge provided by the current mapping results as the inheritance similarity, denoted by $s_I(v, v')$. Formally, $s_I(v, v')$ can be quantified as

$$s_I(v,v') = \begin{cases} \frac{C}{|N_k(v,v')|} \cdot (1 - \frac{|\Delta_v^a - \Delta_{v'}^u|}{\max\{\Delta_v^a, \Delta_{v'}^u\}}) \cdot \sum\limits_{(x,x') \in N_k(v,v')} s(x,x'), & N_k(v,v') \neq \emptyset \\ 0, & \text{otherwise} \end{cases}, \quad (21)$$

where $C \in (0, 1)$ is a constant value representing the *similarity loss exponent*, $N_k(v, v') = (N^a(v) \times N^u(v')) \cap \mathcal{M}_k = \{(x, x') | x \in N^a(v), x' \in N^u(v'), (x, x') \in \mathcal{M}_k\}$ is the set of mapped pairs between $N^a(v)$ and $N^u(v')$ till iteration k, and $s(x, x') \in [0, 1]$ is the overall similarity score between x and x' which is formally defined in the following section.

From the definition of $s_I(v, v')$, we can see that (i) if two nodes have more common neighbors which have been mapped, then their inheritance similarity score is high. For example, in Fig.7, v has more inheritance similarity with v' than with x'. It is reasonable since v and v' are more likely to be the same user in this scenario; (ii) we also count the degree similarity in defining $s_I(v, v')$. If the degree difference between v and v' is small, then a large weight is given to the inheritance similarity; otherwise, a small weight is given; and (iii) we involve the similarity loss in counting $s_I(v, v')$,



(a) St. Andrews vs Facebook (b) Infocom06 vs DBLP (c) Smallblue vs Facebook

Figure 8: Inheritance Similarity.

which implies the inheritance similarity is decreasing with the distance increasing (iteration increasing) between (v, v') and the original seed mappings.

Now, for St. Andrews/Facebook, Infocom06/DBLP, and Smallblue/Facebook, if we assume half of the nodes have been mapped (the first half according to the ID increasing order), then the inheritance similarity between the rest of the nodes in the anonymized graph and the auxiliary graph is shown in Fig.8. From the result, we can observe that under the half number of nodes mapping assumption, some nodes, e.g., nodes 16, 19, 24 in St. Andrews, nodes 40, 56, 64, 72 in Infocom06, and nodes 64, 65, 92, 96, 111, 115 in Smallblue, agree with their counterparts and meanwhile disagree with all the other nodes significantly in the auxiliary graph, which implies that they are potentially easier to be de-anonymized when inheritance similarity is taken as a metric. Note that, in Fig.8, we just randomly assume that the known mapping nodes are the first half nodes in the anonymized and auxiliary graphs. Actually, the accuracy performance of the inheritance similarity measurement could be improved. This is because there are no necessary correlations among the randomly chosen mapping nodes in Fig.8. Nevertheless, in our de-anonymization framework, the obtained mappings in one iteration depend on the mappings in the previous iteration. This strong correlation among mapped nodes allows for the use of the inheritance similarity in practical de-anonymizaiton.

4.5 De-anonymization Algorithm

From the aforementioned discussion, we find that the differentiability of anonymized nodes is different with respect to different similarity measurements. For instance, some nodes have distinctive topological characteristics, e.g., node 16 in the St Andrew dataset, which implies that they can be potentially de-anonymized solely based on the structural similarity. On the other hand, for some nodes, due to the lacking of distinct topological characteristics, the structural similarity based method can only achieve coarse granularity de-anonymization. Nevertheless and fortunately (from the view of adversary), they may become significantly distinguishable with the knowledge of a small amount of auxiliary information, e.g., nodes 14, 19, and 23 in St. Andrews are potentially easy to be de-anonymized based on relative distance similarity. In summary, the analysis on real datasets suggests to us to define a unified measurement to properly involve multiple similarity metrics for effective de-anonymization. To this end, we define a Unified Similarity (US) measurement by considering the structural similarity, relative distance similarity, and inheritance similarity synthetically for $v \in \Lambda^\delta(M_k^a)$ and $v' \in \Lambda^\delta(M_k^u)$ in the k-th iteration of our de-anonymization framework as

$$s(v, v') = c_S \cdot s_S(v, v') + c_D \cdot s_D(v, v') + c_I \cdot s_I(v, v'), \tag{22}$$

where $c_S, c_D, c_I \in [0, 1]$ are constant values indicating the weights of structural similarity, relative distance similarity, and inheritance similarity, respectively, and $c_S + c_D + c_I = 1$. In addition, we define $s(v, v') = 1$ if $(v, v') \in \mathcal{M}_s$. Now, we are ready to present our US based **D**e-**A**nonymization framework, denoted by DA, which is shown in Algorithm 1.

In Algorithm 1, $B_k = (\Lambda^\delta(M_k^a) \cup \Lambda^\delta(M_k^u), E_k^b, W_k^b)$ is a weighted bipartite graph defined on the intended de-anonymizing nodes during the k-th iteration, where $E_k^b = \{l_{v,v'}^b | \forall v \in \Lambda^\delta(M_k^a), \forall v' \in \Lambda^\delta(M_k^u)\}$, and $W_k^b = \{w_{v,v'}^b\}$ is the set of all the possible weights on the links in E_k^b. Here, for $\forall (v, v') \in E_k^b$, the weight on this link is defined as

Algorithm 1: US based **D**e-**A**nonymization (DA) framework

```
input  : $G^a, G^u, \mathcal{M}_s$
output: de-anonymization of $G^a$
$\mathcal{M}_0 = \mathcal{M}_s$, $k = 0$, $flag$ = true;
while $flag$ = true do
    calculate $\Lambda^\delta(M_k^a)$ and $\Lambda^\delta(M_k^u)$;
    if $\Lambda^\delta(M_k^a) = \emptyset$ or $\Lambda^\delta(M_k^u) = \emptyset$ then
        output $\mathcal{M}_k$ and break;
    for every $v \in \Lambda^\delta(M_k^a)$ do
        for every $v' \in \Lambda^\delta(M_k^u)$ do
            calculate $s(v, v')$;
    construct a weighted bipartite graph $B_k = (\Lambda^\delta(M_k^a) \cup \Lambda^\delta(M_k^u), E_k^b, W_k^b)$
    between nodes $\Lambda^\delta(M_k^a)$ and $\Lambda^\delta(M_k^u)$ based on $s(v, v')$;
    use the Hungarian algorithm to obtain a maximum weighted bipartite
    matching of $B_k$, denoted by $\mathcal{M}' = \{(v, v') | v \in \Lambda^\delta(M_k^a), v' \in \Lambda^\delta(M_k^u)\}$;
    for every $(x, x') \in \mathcal{M}'$ do
        if $s(x, x') < \theta$ then
            $\mathcal{M}' = \mathcal{M}' \setminus \{(x, x')\}$;
    let $K = \max\{1, \lceil |\epsilon \cdot \mathcal{M}'| \rceil\}$ and for $\forall (x, x') \in \mathcal{M}'$, if $s(x, x')$ is not the
    Top-$K$ mapping score in $\mathcal{M}'$ then
        $\mathcal{M}' = \mathcal{M}' \setminus \{(x, x')\}$, i.e. only keep the Top-$K$ mapping pairs in $\mathcal{M}'$;
    if $\mathcal{M}' = \emptyset$ then
        output $\mathcal{M}_k$ and break;
    $\mathcal{M}_{k+1} = \mathcal{M}_k \cup \mathcal{M}'$;
    $k$++;
```

the US score between the associated two nodes, i.e., $w^b_{v,v'} = s(v, v')$. Parameter θ is a constant value named de-anonymization threshold to decide whether a node mapping is accepted or not. Parameter $\epsilon \in (0, 1]$ is the mapping control factor, which is used to limit the maximum number of mappings generated during each iteration. By ϵ, even if there are many mappings with similarity score greater than the de-anonymization threshold, we only keep the $K = \max\{1, \lceil|\epsilon \cdot \mathcal{M}'|\rceil\}$ more confident mappings.

We give further explanation on the idea of Algorithm DA as follows. The de-anonymization is bootstrapped with an initial seed mapping (line 1) and starts the iteration procedure (lines 2-19). During each iteration, the intended de-anonymizing nodes are calculated first based on the mappings obtained in the previous iteration (lines 3-5) followed by calculating the US scores between nodes in $\Lambda^\delta(M^a_k)$ and nodes in $\Lambda^\delta(M^u_k)$ (lines 6-8). Subsequently, based on the obtained US scores, a weighted bipartite graph is constructed between nodes in $\Lambda^\delta(M^a_k)$ and nodes in $\Lambda^\delta(M^u_k)$ (line 9). Then, we compute a maximum weighted bipartite matching $\mathcal{M}'$ on the constructed bipartite graph by the Hungarian algorithm (line 10). To improve the de-anonymization accuracy, we apply two important rules to refine $\mathcal{M}'$: (*i*) by defining a de-anonymization threshold θ, we eliminate the mappings with low US scores in $\mathcal{M}'$ (lines 11-13). This is because we are not confident to take the mappings with low US scores ($< \theta$) as correct de-anonymizaiton, and more improtantly, they may be more accurately de-anonymized in the following iterations by utilizing confident mapping information obtained in this iteration (this can be achieved since we involve inheritance similarity in the US definition); and (*ii*) we introduce a mapping control factor ϵ, or K equivalently, to limit the maximum number of mappings been recognized as correct de-anonymization (lines 14-15). During each iteration, only K mappings with highest US scores will be taken as correct de-anonymization with confidence even if more mappings having US scores greater than the de-anonymizaiton threshold. This strategy has two benefits. On one hand, only highly confident mappings are kept, which could improve the de-anonymization accuracy. On the other hand, for the mappings been rejected, again, they may be better re-de-anonymized in the following iterations by utilizing the more confident knowledge of the Top-K mappings from this iteration (lines 18-19).

CHAPTER V

GENERALIZED SCALABLE DE-ANONYMIZATION

In this chapter, we extend DA to more general scenarios such as large-scale data de-anonymization including the situation that the anonymized graph and the auxiliary graph are partially overlapped, and disconnected anonymized graphs or auxiliary graphs.

5.1 De-anonymization on Large-Scale Datasets

The proliferation of social applications and services has resulted in the production of significant amounts of data. To de-anonymize large-scale data, besides the de-anonymization accuracy, efficiency and scalability are also important concerns. Another predicament in practical de-anonymization, which is omitted in existing de-anonymization attacks, is that we do not actually know how large the overlap between the anonymized data and the auxiliary data even we have a lot of auxiliary information available. Therefore, it is unadvisable to do de-anonymization based on the entire anonymized and auxiliary graphs directly, which might cause low de-anonymization accuracy as well as high computational overhead.

To address the aforementioned predicament, guarantee the accuracy of DA, and simultaneously improve de-anonymization efficiency and scalability on large-scale data, we extend DA to an **A**daptive **D**e-**A**nonymization framework, denoted by ADA. ADA adaptively de-anonymizes G^a starting from a Core Matching Subgraph (CM-S), which is formally defined as follows. Let $\mathcal{M}_s$ be the initial seed mapping between the anonymized graph G^a and the auxiliary graph G^u. Furthermore, define $V_s^a = \bigcup\limits_{x,y \in M_0^a} \{v | v \text{ lies on } p^a(x,y) \in \mathbb{P}^a(x,y)\}$, i.e., V_s^a is the union of all the nodes on the shortest paths among all the seeds in G^a, and $V_c^a = V_s^a \cup \Lambda^\delta(V_s^a)$, i.e., V_c^a

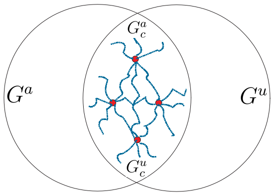


Figure 9: Core Matching Subgraph (CMS). Initial seed mappings are denoted by red nodes.

is the union of V_s^a and the δ-hop mapping spanning set of V_s^a. Then, we define the initial CMS on G^a as the subgraph of G^a on V_c^a, i.e., $G_c^a = G^a[V_c^a]$ (as shown in Fig.9). Similarly, we can define $V_s^u = \bigcup\limits_{x',y' \in M_0^u} \{v' | v' \text{ lies on } p^u(x',y') \in \mathbb{P}^u(x',y')\}$ and $V_c^u = V_s^u \cup \Lambda^\delta(V_s^u)$. Then, the initial CMS on G^u is $G_c^u = G^a[V_c^u]$ (as shown in Fig.9).

The CMS is generally defined for two purposes. On one hand, we can employ a CMS to adaptively and roughly estimate the overlap between G^a and G^u as shown in Fig.9 in terms of the seed mapping information. On the other hand, we propose to start the de-anonymization from the CMSs in G^a and G^u, by which the de-anonymization is smartly limited to start from two small subgraphs with more information confidence, and thus we could improve the de-anonymization accuracy and reduce the computational overhead.

Algorithm 2: Adaptive **De**-**A**nonymization (ADA)

```
input  : $G^a, G^u, \mathcal{M}_s$
output: de-anonymization of $G^a$
1 generate $G_c^a$ and $G_c^u$ from $G^u$ and $G^a$ respectively;
2 run DA for $G_c^a$ and $G_c^u$;
3 if Step 2 is ended on the condition that $\Lambda^\delta(M_k^a) = \emptyset$ or $\Lambda^\delta(M_k^u) = \emptyset$ then
4     if $\Lambda^\mu(V_c^a) = \emptyset$ or $\Lambda^\mu(V_c^u) = \emptyset$ then
5         return;
6     $V_c^a = V_c^a \cup \Lambda^\mu(V_c^a)$, $V_c^a = V_c^a \cup \Lambda^\mu(V_c^a)$;
7     $G_c^a = G^a[V_c^a]$, $G_c^u = G^u[V_c^u]$;
8     go to Step 2 to de-anonymize unmapped nodes in updated $G_c^a$ and $G_c^u$;
```

Now, based on CMS, we discuss ADA as shown in Algorithm 2. In Algorithm 2, μ is the adaptive factor which controls the spanning size of the CMS during each adaptive iteration. The basic idea of ADA is as follows. We start the de-anonymization from CMSs G_c^a and G_c^u by running DA (lines 1-2). If DA is ended with $\Lambda^\delta(M_k^a) = \emptyset$ or $\Lambda^\delta(M_k^u) = \emptyset$, then the actual overlap between G^a and G^u might be larger than G_c^a/G_c^u since more nodes could be mapped. Therefore, we enlarge the previous considering CMS G_c^a/G_c^u by involving more nodes $\Lambda^\mu(V_c^a)/\Lambda^\mu(V_c^u)$ and repeat the de-anonymization for unmapped nodes in the updated G_c^a/G_c^u (lines 3-8).

5.2 Disconnected Datasets

In reality, when we employ a graph G^a/G^u to model the anonymized/auxiliary data, G^a/G^u might be not connected. In this case, G^a and G^u can be represented by the union of connected components as $\bigcup\limits_i G_i^a$ and $\bigcup\limits_j G_j^u$ respectively, where G_i^a and G_j^u are some connected components. Now, when defining the structural similarity, relative distance similarity, or inheritance similarity, we change the context from G^a/G^u to components G_i^a/G_j^u. Then, we can apply DA/ADA to conduct de-anonymization.

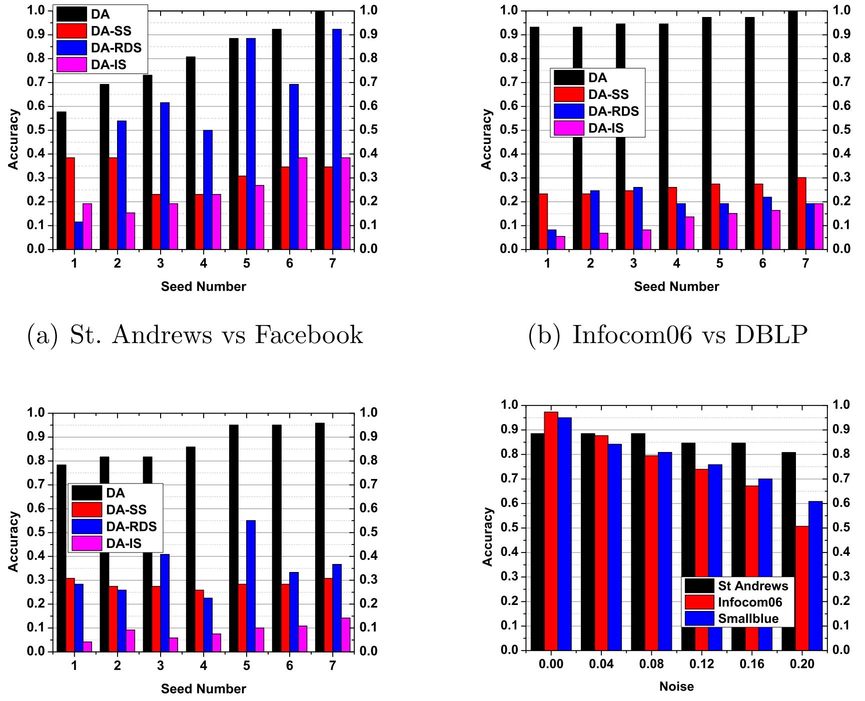
CHAPTER VI

EXPERIMENTS

In this chapter, we examine the performance of the presented de-anonymization attack on real datasets. Particularly, we will validate DA/ADA on mobility traces (St. Andrews/Facebook, Infocom06/DBLP, Smallblue/Facebook), weighted data (Arnet-Miner), as well as social network data (Google+, Facebook).

6.1 De-anonymizing Mobility Traces

By utilizing the corresponding social networks as auxiliary information, we employ the presented de-anonymization algorithm DA to de-anonymize the three well known mobility traces St. Andrews, Infocom06, and Smallblue. The results are shown in Fig.10 (a)-(c), where DA denotes the presented US-based de-anonymization framework, and DA-SS, DA-RDS, and DA-IS represent the de-anonymization based on structural similarity solely (by setting $c_S = 1$ and $c_D = c_I = 0$ in US), relative distance similarity solely (by setting $c_D = 1$ and $c_S = c_I = 0$ in US), and inheritance similarity solely (by setting $c_I = 1$ and $c_S = c_D = 0$ in US), respectively. From Fig.10 (a)-(c), we can see that (*i*) the presented de-aonymization framework is very effective even with a small amount of auxiliary information. For instance, DA can successfully de-anonymize 93.2% of the Infocom06 data just with the knowledge of one seed mapping. For St. Andrews and Smallblue, DA can also achieve accuracy of 57.7% and 78.3% respectively with one seed mapping. Furthermore, DA can successfully de-anonymize all the data in St. Andrews and Smallblue and 96% of the data of Smallblue with the knowledge of 7 seed mappings; and (*ii*) the US-based de-anonymization is much more effective and stable than structural, relative distance, or inheritance similarity solely based de-anonymization. The reason is that US tries to



(a) St. Andrews vs Facebook

(b) Infocom06 vs DBLP

(c) Smallblue vs Facebook

(d) De-anonymization with noise

Figure 10: De-anonymization performance versus Similarity measurements, seed selection, and noise. Default parameter settings: $C = 0.9$, $c_S = 0.1$, $c_D = 0.8$, $c_I = 0.1$, $\theta = 0.6$, $\delta \in \{1, 2\}$, $\epsilon = 0.5$, and seed number is 5.

distinguish a node from multiple perspectives, which is more efficient and comprehensive. As the analysis shown in Chapter 4, the nodes can be easily differentiated with respect to one measurement but might be indistinguishable with respect to another measurement. Consequently, synthetically characterizing a node as in US is more powerful and stable.

We also examine the robustness of the presented de-anonymization attack to noise and the result is shown in Fig.10 (d) (on the knowledge of 5 seed mappings). In the experiment, we only add noise to the anonymized data. According to the same argument in [18], the noise in the auxiliary data can be counted as noise in the anonymized data. To add p percent of noise to the anonymized data, we randomly add $\frac{p}{2} \cdot |E^a|$ spurious connections to and meanwhile delete $\frac{p}{2} \cdot |E^a|$ existing connections from the anonymized graph (a node may become *isolated* after the noise adding process). For instance, in Fig.10 (d), 20% of noise implies we add 10% spurious connections and delete 10% existing connections of $|E^a|$ from the anonymized data. From Fig.10 (d), we can see that the presented de-anonymization framework is robust to noise. Even if we change 20% of the connections in the anonymized data, the achieved accuracies on St. Andrews, Infocom06, and Smallblue are still 80.8%, 50.7%, and 60.8%, respectively. Note that, when 20% of the connections have been changed, the structure of the anonymized data is significantly changed. In practical, if the anonymized data release is initially for research purposes, e.g., data mining, this structural change may make the data useless. However, by considering multiple perspectives to distinguish a node, the anonymized data can still be de-anonymized as shown in Fig.10 (d), which confirms the assertion in [18] that structure change may not provide effective privacy protection.

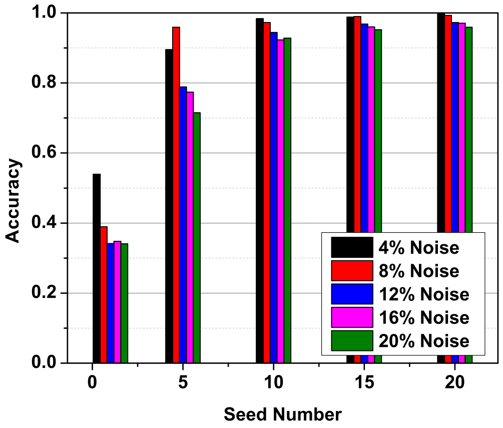
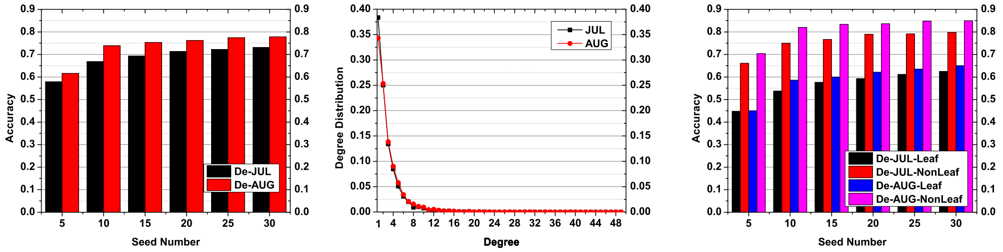


Figure 11: De-anonymize ArnetMiner. Default parameter settings: $\alpha = 1.5$, $C = 0.9$, $c_S = 0.2$, $c_D = 0.6$, $c_I = 0.2$, $\theta = 0.6$, $\delta \in \{1, 2\}$, $\mu \in \{1, 2, 3\}$, and $\epsilon = 0.5$.

6.2 De-anonymizing ArnetMiner

ArnetMiner is a coauthor dataset consisting of 1,127 authors and 6,690 coauthor relationships. Consequently, ArnetMiner can be modeled by a weighted graph where the weight on each relationship indicates the number of coauthored papers by the two authors. To examine the de-anonymization framework, we first anaonymize ArnetMiner by adding p percent noise as explained in the previous section. Furthermore, for each added spurious coauthor relationship, we also randomly generate a weight in $[1, A_{\max}]$, where $A_{\max}$ is the maximum weight in the original ArnetMiner graph. Then, we de-anonymize the anonymized data using the original ArnetMiner data and the result is shown in Fig.11.

From Fig.11, we can observe that the presented de-anonymization framework is very effective on weighted data. With only knowledge of one seed mapping, more than a half (53.9%) and one-third (34.1%) of the authors can be de-anonymized even with noise levels of 4% and 20%, respectively. Furthermore, when adding 20% of noise to the anonymized data, the presented de-anonymization framework achieves 71.5% accuracy if 5 seed mappings are available and 92.8% accuracy if 10 seed mappings are available; (ii) the presented de-anonymization framework is robust to noise on



(a) De-anonymize Google+

(b) Degree distribution of Google+

(c) Degree distribution of Google+

Figure 12: De-anonymize Google+. Default parameter settings: $C = 0.9$, $c_S = 0.2$, $c_D = 0.6$, $c_I = 0.2$, $\theta = 0.6$, $\delta \in \{1, 2\}$, $\mu \in \{1, 2, 3\}$, and $\epsilon = 0.5$.

weighted data. When we have 10 or more seed mappings, the accuracy degradation of our de-anonymization algorithm is small even with more noise, e.g., the accuracy is degraded from 99.7% in the 4%-noise case to 96% in the 20%-noise case; and (*iii*) if the available number of seed mappings is 10, the knowledge brought by more seed mappings cannot improve the de-anonymization accuracy significantly. This is because the achieved accuracy on the knowledge of 10 seed mappings is already about 95%. Therefore, to de-anonymize a data set, it is not necessary to spend efforts to obtain a lot of seed mappings. As in this case, to de-anonymize most of the authors, 5 to 10 seed mappings are sufficient.

6.2.1 De-anonymizing Google+

Now, we validate the presented de-anonymization framework on the two Google+ datasets JUL (5,200 users and 7,062 connections) and AUG (5,200 users and 7,813 connections). We first utilize AUG as auxiliary data to de-anonymize JUL denoted by De-JUL, i.e., use future data to de-anonymize historical data, and then utilize JUL to de-anonymize AUG denoted by De-AUG, i.e., use historical data to de-anonymize future data. The results are shown in Fig.12 (a). Again, from Fig.12 (a), we can see that the presented de-anonymization framework is very effective. Just based on the knowledge of 5 seed mappings, 57.9% of the users in JUL and 61.6% of the users in

AUG can be successfully deanonymized. When 10 seed mappings are available, the de-anonymization accuracy can be improved to 66.8% on JUL and 73.9% on AUG, respectively.

However, we also have two other interesting observations from Fig.12 (a): (*i*) when the number of available seed mappings is above 10, the performance improvement is not as significant as on previous datasets (e.g., mobility traces, ArnetMiner) even the de-anonymization accuracy is around 70% for JUL and 75% for AUG; and (*ii*) De-AUG has a better accuracy than De-JUL, which implies that the AUG dataset is easier to de-anonymize than the JUL dataset. To explain the two observations, we assert this is because of the structural property of the two datasets. Follow this direction, we investigate the degree distribution of JUL and AUG as shown in Fig.12 (b). From Fig.12 (b), we can see that the degree of both JUL and AUG generally follows a *heavy-tailed distribution*. In particular, 38.4% of the users in JUL and 34.3% of the users in AUG have degree of one, named *leaf users*. This is normal since Google+ was launched in early July 2011, and JUL and AUG are datasets crawled in July and August of 2011, respectively. That is also why JUL has more leaf users than AUG (a user connects more people later). Now, we argue that the leaf users cause the difficulty in improving the de-anonymization accuracy. From the perspective of graph theory, the leaf users limit not only the performance of our de-anonymization framework but also the performance of any de-anonymization algorithm. An explanatory example is as follows. Suppose $v \in V^a$ is successfully de-anonymized to $v' \in V^u$. In addition, the two neighbors x and y of v and the two neighbors x' and y' of v' are all leaf users. Then, even $x' = \gamma(x)$, $y' = \gamma(y)$, and v has been successfully de-anonymized to v', it is still difficult to make a decision to map x (or y) to x' or y' since $s(x, x') \approx s(x, y')$ from the view of graph theory. Consequently, to accurately distinguish x, further knowledge is required.

To support our argument, we take an insightful look on the experimental results.

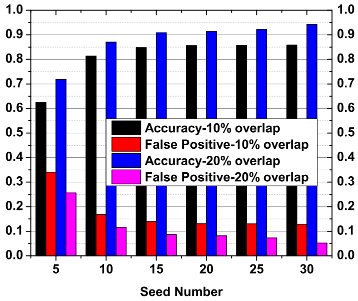


Figure 13: De-anonymize Facebook. Default parameter settings: $C = 0.9$, $c_S = 0.2$, $c_D = 0.6$, $c_I = 0.2$, $\theta = 0.8$, $\delta \in \{1, 2\}$, $\mu \in \{1, 2, 3\}$, and $\epsilon = 0.5$.

For each successfully de-anonymized user in JUL and AUG, we classify the user in terms of its degree into one of two sets: *leaf user set* if its degree is one or *non-leaf user set* if its degree is greater than one. Then, we re-calculate the de-anonymization accuracy for leaf users and non-leaf users and the results are shown in Fig.12 (c), where De-JUL-Leaf/De-AUG-Leaf represents the ratio of leaf nodes that have been successfully de-anonymized in JUL/AUG while De-JUL-NonLeaf/De-AUG-NonLeaf represents the ratio of non-leaf users that have been successfully de-anonymized in JUL/AUG. From Fig.12 (c), we can see that (*i*) the successful de-anonymization ratio on non-leaf users is higher than that on leaf users in JUL and AUG. This is because non-leaf users carry more structural information; and (*ii*) considering the results shown in Fig.12 (a), the de-anonymization accuracy on non-leaf users is higher than the overall accuracy and the de-anonymization accuracy on leaf users is lower than the overall accuracy. The two observations on Fig.12 (c) confirms our argument that leaf users are more difficult than non-leaf users to de-anonymize. Furthermore, this is also why De-AUG has higher accuracy than De-JUL in Fig.12 (a). AUG is easier to de-anonymize since it has less leaf users than JUL.

6.3 De-anonymizing Facebook

Finally, we examine ADA on a Facebook dataset, which consists of 63,731 users and 1,269,502 "friendship" users. Based on the hand labeled ground truth, we partition the datasets into two about-equal parts utilizing the method employed in [18], and then we take one part as auxiliary data to de-anonymize the other part. When the two parts only have 10% and 20% users in common (i.e., only 10% and 20% overlap between the anonymized graph and the auxiliary graph), the achievable accuracy and the induced false positive error of ADA are shown in Fig.13. As a fact, most of the existing de-anonymization attacks are not very effective for the scenario that the overlap between the anaonymized data and the auxiliary data is small or even cannot work totally. Surprisingly, for ADA, we can observe from Fig.13 that (*i*) based on the proposed CMS, ADA can successfully de-anonymize 62.4% of the common users with false positive error of 34.1% when the overlap is 10% and 71.8% of the common users with false positive error of 25.6% when the overlap is 20% with the knowledge of just 5 seed mappings; (*ii*) the de-anonymization accuracy is improved to 81.3% (respectively, 85.6%) and the false positive error is decreased to 16.8% (respectively, 13%) when the overlap is 10% and 10 (respectively, 20) seed mappings available, and the de-anonymization accuracy is improved to 87% (respectively, 90.8%) and the false positive error is decreased to 11.6% (respectively, 8.6%) when the overlap is 20% and 10 (respectively, 20) seed mappings available, which demonstrate that ADA is very effective in dealing with the partial data overlap situation; and (*iii*) ADA has a higher de-anonymization accuracy and lower false positive error in the 20% data overlap scenario than that in the 10% data overlap scenario. This is because a larger overlap size implies a common node will carry much more similar structural information in both graphs. From Fig.13, we can also see that 10 seed mappings are sufficient to achieve high de-anonymization accuracy and low false positive error. Therefore, ADA is applicable with efficiency and performance guarantee in practical.

CHAPTER VII

CONCLUSION

In this thesis, we present a novel, robust, and effective de-anonymization attack to both mobility trace data and social network data. First, we design three de-anonymization metrics which take into account both local and global structural characteristics of data, the information obtained from auxiliary data, as well as the knowledge inherited from on-going de-anonymization results. When analyzing the three metrics on real datasets, we find that some data can potentially be de-anonymized accurately and the other can be de-anonymized with coarse granularity. Subsequently, we introduce a Unified Similarity (US) measurement which synthetically incorporates the three defined metrics. Based on US, we propose a De-Anonymization (DA) framework, which iteratively de-anonymizes data with accuracy guarantee. Then, to de-anonymize large scale data without the knowledge on the overlap size between the anonymized data and the auxiliary data, we generalize DA to an Adaptive De-Anonymization (ADA) framework. ADA works on two Core Matching Subgraphs (CMSs) adaptively, by which the de-anonymization is limited to the overlap area of the anonymized data and the auxiliary data, followed by improving de-anonymization accuracy and reducing computational overhead. Finally, we apply the presented de-anonymization attack to three mobility trace data sets: St. Andrews/Facebook, Infocom06/DBLP, and Smallblue/Facebook, and three relatively large social network datasets: ArnetMiner (weighted data), Google+, and Facebook. The experimental results demonstrate that the presented de-anonymization framework is very effective and robust to noise.

REFERENCES

[1] Alvisi, L., Clement, A., Epasto, A., Lattanzi, S., and Panconesi, A., "Sok: The evolution of sybil defense via social networks," *S&P*, 2013.

[2] Backstrom, L., Dwork, C., and Kleinberg, J., "Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography," *WWW*, 2007.

[3] Bigwood, G., Rehunathan, D., Bateman, M., Henderson, T., and B-hatti, S., "Crawdad data set st_andrews/sassy (v. 2011-06-03)," 2011.

[4] Campan, A. and Truta, T. M., "A clustering approach for data and structural anonymity in social networks," *PinKDD*, 2008.

[5] CRAWDAD 2014.

[6] Egele, M., Kruegel, C., Kirda, E., and Vigna, G., "Pios: Detecting privacy leaks in ios applications," *NDSS*, 2011.

[7] Gong, N. Z., Talwalkar, A., Mackey, L., Huang, L., Shin, E. C. R., Stefanov, E., Shi, E., and Song, D., "Jointly predicting links and inferring attributes using a social-attribute network (san)," *SNA-KDD*, 2012.

[8] Gong, N. Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., and Song, D., "Evolution of social-attribute networks: Measurements, modeling, and implications using google+," *IMC*, 2012.

[9] Hay, M., Miklau, G., Jensen, D., Towsley, D., and Weis, P., "Resisting structural re-identification in anonymized social networks," *VLDB*, 2008.

[10] Hornyack, P., Han, S., Jung, J., Schechter, S., and Wetherall, D., "These aren't the droids you're looking for: Retrofitting android to protect data from imperious applications," *CCS*, 2011.

[11] Ji, S., Li, W., Mittal, P., Hu, X., and Beyah, R., "Structural data de-anonymization: Quantification, practice, and implications," *CCS*, 2014.

[12] Ji, S., Li, W., Mittal, P., Hu, X., and Beyah, R., "On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge," *NDSS*, 2015.

[13] Ji, S., Li, W., Mittal, P., Hu, X., and Beyah, R., "Secgraph: A uniform and open-source evaluation system for graph data anonymization and de-anonymization," *USENIX Security*, 2015.

[14] Kruskal, J. and Wish, M., "Multidimensional scaling," *Sage Publication*, 1978.

[15] Liu, K. and Terzi, E., "Towards identity anonymization on graphs," *SIGMOD*, 2008.

[16] McAuley, J. and Leskovec, J., "Learning to discover social circles in ego networks," *NIPS*, 2012.

[17] Narayanan, A. and Shmatikov, V., "Robust de-anonymization of large sparse datasets (de-anonymizing the netflix prize dataset)," *S&P*, 2008.

[18] Narayanan, A. and Shmatikov, V., "De-anonymizing social networks," *S&P*, 2009.

[19] Opsahl, T., Agneessens, F., and Skvoretz, J., "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks*, vol. 32, pp. 245–251, 2010.

[20] S. Nilizadeh, A. K. and Ahn, Y.-Y., "Community-enhanced de-anonymization of online social networks," *CCS*, 2014.

[21] Scott, J., Gass, R., Crowcroft, J., Hui, P., Diot, C., and Chaintreau, A., "Crawdad data set cambridge/haggle (v. 2009-05-29)," May 2009.

[22] SecGraph 2015.

[23] Singh, K., Bhola, S., and Lee, W., "xbook: Redesigning privacy control in social networking platforms," *USENIX*, 2009.

[24] Smallblue 2009.

[25] SNAP 2014.

[26] Srivatsa, M. and Hicks, M., "Deanonymizing mobility traces: Using social networks as a side-channel," *CCS*, 2012.

[27] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z., "Arnetminer: Extraction and mining of academic social networks," *KDD*, 2008.

[28] Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P., "On the evolution of user interaction in facebook," *WOSN*, 2009.

[29] Yu, H., Gibbons, P. B., Kaminsky, M., and Xiao, F., "Sybillimit: A near-optimal social network defense against sybil attacks," *S&P*, 2008.

[30] Yu, H., Kaminsky, M., Gibbons, P. B., and Flaxman, A. D., "Sybilguard: Defending against sybil attacks via social networks," *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, pp. 576–589, 2008.

[31] Yu, H., Shi, C., Kaminsky, M., Gibbons, P. B., and Xiao, F., "Dsybil: Optimal sybil-resistance for recommendation systems," *S&P*, 2009.

[32] Zheleva, E. and Getoor, L., "Preserving the privacy of sensitive relationships in graph data," *PinKDD*, 2007.