General Graph Data De-Anonymization: From Mobility Traces to Social Networks

SHOULING JI and WEIQING LI, Georgia Institute of Technology
MUDHAKAR SRIVATSA, IBM T. J. Watson Research Center
JING SELENA HE, Kennesaw State University
RAHEEM BEYAH, Georgia Institute of Technology

When people utilize social applications and services, their privacy suffers a potential serious threat. In this article, we present a novel, robust, and effective de-anonymization attack to mobility trace data and social data. First, we design a Unified Similarity (US) measurement, which takes account of local and global structural characteristics of data, information obtained from auxiliary data, and knowledge inherited from ongoing de-anonymization results. By analyzing the measurement on real datasets, we find that some data can potentially be de-anonymized accurately and the other 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 accuracy guarantee. Then, to de-anonymize large-scale data without knowledge of the overlap size between the anonymized data and the auxiliary data, we generalize DA to an Adaptive De-Anonymization (ADA) framework. By smartly 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 datasets: ArnetMiner, Google+, and Facebook. The experimental results demonstrate that the presented de-anonymization framework is very effective and robust to noise.

The source code and employed datasets are now publicly available at SecGraph [2015].

CCS Concepts: • **Security and privacy → Pseudonymity, anonymity and untraceability**; • **Security and privacy** → *Data anonymization and sanitization*

Additional Key Words and Phrases: Graph de-anonymization, social networks, mobility traces

This work was partly supported by the National Science and Technology Support Program of China (2014BAH24F01). Mudhakar Srivatsa's research was partly sponsored by the US Army Research Laboratory and the UK Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory, the U.S. Government, the UK Ministry of Defense, or the UK Government. The U.S. and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. Jing S. He's research is partly supported by the Kennesaw State University College of Science and Mathematics Interdisciplinary Research Opportunities (IDROP) Program.

Authors' addresses: S. Ji is with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China and with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA; email: sji@gatech.edu; W. Li is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA; email: wli64@gatech.edu; M. Srivatsa is with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA; email: msrivats@us.ibm.com; J. S. He is with the Department of Computer Science, Kennesaw State University, Marietta, GA 30060, USA; email: jhe4@kennesaw.edu; R. Beyah is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA; email: rbeyah@ece.gatech.edu.

DOI: http://dx.doi.org/10.1145/2894760

ACM Reference Format:
Shouling Ji, Weiqing Li, Mudhakar Srivatsa, Jing Selena He, and Raheem Beyah. 2016. General graph data de-anonymization: From mobility traces to social networks. ACM Trans. Inf. Syst. Secur. 18, 4, Article 12 (April 2016), 29 pages.
DOI: http://dx.doi.org/10.1145/2894760

1. INTRODUCTION

Social networking services are a fast-growing business nowadays. The development of smartphone 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 society, social networks and data analysis have attracted more and more research interest [Backstrom et al. 2007; Narayanan and Shmatikov 2009; Srivatsa and Hicks 2012; Ji et al. 2014, 2015a, 2015b].

When users participate in online social network activities (e.g., create personal portfolios and connect to social friends) or utilize social network functions (e.g., post current location or share information with virtual social friends), people's privacy suffers a potential serious threat. On the other hand, to utilize the huge amount of users' data for commercial or academic purposes, social network owners usually release social data for research (data mining) or transfer data to business partners for target advertising [Narayanan and Shmatikov 2009]. Furthermore, the mighty advance of mobile computing and communication technology enables mobile devices such as smartphones to gather abundant information about users [Srivatsa and Hicks 2012]. For example, users can update their location position or show their sharing/following information through Twitter/Facebook on smartphones easily.

To protect users' privacy, social network owners and services providers usually anonymize data by removing "Personally Identifiable Information (PII)" before releasing the data to the public. However, in reality, this data anonymization is vulnerable to a new *social auxiliary information-based data de-anonymization attack* [Backstrom et al. 2007; Narayanan and Shmatikov 2009; Srivatsa and Hicks 2012]. The *practicality* and *effectiveness* come from two fundamental facts. First, when network owners and services providers publish data, only naive anonymization techniques are applied to remove basic PII. For example, even for the carefully processed (by unknown sampling techniques) and anonymized (incorporating data perturbation) Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix and which was released for the research contest purpose, its structure itself carries enough information for effective privacy breaching [Narayanan and Shmatikov 2008]. Furthermore, existing anonymization techniques (e.g., Campan and Truta [2008], Hay et al. [2008], Liu and Terzi [2008], and Zheleva and Getoor [2007]) have several limitations, such as making impractical assumptions about social data or knowledge of adversaries, not being scalable, and so forth (detailed limitation analysis is shown in the related work section), which prevent them from being workable in reality [Backstrom et al. 2007; Narayanan and Shmatikov 2009; Srivatsa and Hicks 2012]. The second fact is the wide and common availability of social auxiliary information [Backstrom et al. 2007; Narayanan and Shmatikov 2009; Srivatsa and Hicks 2012]. As indicated in Narayanan and Shmatikov [2009] and Srivatsa and Hicks [2012], adversaries can obtain social auxiliary information easily or with a few efforts through multiple channels, for example, academic and government data mining, advertising, third-party applications, data aggregation and inferring, privacy attack and acquiring, and 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 successful privacy breaching.

A few de-anonymization attacks have been designed for social data [Backstrom et al. 2007; Narayanan and Shmatikov 2009] or mobility trace data [Srivatsa and Hicks 2012]. However, existing works are limited due to one or several reasons, for example, scalability, generality, and robustness. Our work improves existing works in some or all of the following aspects. First, we significantly improve the de-anonymization accuracy and decrease the computational complexity by proposing a novel *Core Matching Subgraphs* (CMSs)-based adaptive de-anonymization strategy. Second, besides utilizing a node's local property, we incorporate a 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 data, directed and undirected data graphs, and weighted and unweighted datasets. We give the detailed analysis and remarks in the related work section (Section 2).

In summary, our main contributions in this article 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.[1] 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 that have already been de-anonymized. We also examine how the three measurements function on real datasets. By 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 De-Anonymization (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 knowledge of the overlap size between the anonymized data and the auxiliary data, we generalize DA to an Adaptive De-Anonymization (ADA) framework. ADA adaptively conducts data de-anonymization starting from two *Core Matching Subgraphs*, which are defined to estimate the overlap size between the anonymized data and the auxiliary data. By smartly 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 [Bigwood et al. 2011], Infocom06 [Scott et al. 2009], and Smallblue [Smallblue 2009]. The experimental results demonstrate that the presented de-anonymization attack is very effective and robust. With only the knowledge of

[1] In this article, we use topological characteristics and structure/structural characteristics interchangeability.

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, respectively (with five seed mappings).

(5) We also examine the presented de-anonymization attack on social datasets: ArnetMiner (a weighted coauthor dataset consists of 1,127 authors and 6,690 coauthor relationships) Google+ (two datasets with one consisting of 5,200 users and 7,062 connections and the other consisting 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, and 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 article is organized as follows. In Section 2, we survey the most related work. In Section 3, we give the preliminaries and considered data model. In Section 4, the de-anonymization framework is presented. In Section 5, the proposed de-anonymization framework is refined and extended to general large-scale social datasets. We illustrate and discuss the results from extensive experiments on real social and mobility datasets in Section 6. Finally, we conclude this article in Section 7.

2. RELATED WORK

In this section, we survey the related work. We first review the network science metrics that we leverage to measure the similarity, especially the structural similarity, between anonymized users and auxiliary users. Then, we survey specific anonymization techniques for and de-anonymization attacks on social and mobility datasets. Finally, we discuss the differences that distinguish the proposed de-anonymization attack from existing de-anonymization attacks.

2.1. Metrics and Similarity

In the design of our de-anonymization attack, one crucial step is to measure the structural similarity between an anonymized user and an auxiliary user (that can be considered as a known user, formally defined in Section 3). When measuring the structural similarity between two nodes (users), we employ several metrics from the network science. We briefly introduce these metrics in this subsection.

In graph theory and network science, *centrality measurements* are widely employed to measure the importance of a node within a graph/network [Centrality 2015; Newman 2010]. In this article, we use *degree centrality*, *closeness centrality*, and *betweenness centrality*, as well as their weighted versions, to measure the structural properties of a node within a graph (formal definitions are given in Section 4). Degree centrality measures the number of edges connected to a node, which shows the local structural characteristic (importance) of a node within a graph [Freeeman 1978; Newman 2010]. Closeness centrality measures the mean distance from a node to other nodes in a graph (component in a disconnected graph), which is one widely employed global structural characteristic of a node [Newman 2010]. It was first introduced by Bavelas [1950] (in the name of *reciprocal of the farness*). Later, it was rephrased by Freeeman [1978].

[Rochat 2009] extended the closeness centrality to the scenario of disconnected graphs. In another independent work, Opsahl [2010] also extended the concept of closeness centrality to disconnected graphs. Betweenness centrality measures the extent to which a node lies on the shortest paths between other nodes, which is another widely employed global structural characteristic of a node [Freeeman 1978; Newman 2010]. Its formalization is usually attributed to Freeman in Freeeman [1978]. Recently, Opsahl et al. [2010] generalized the definitions of degree centrality, closeness centrality, and betweenness centrality to weighted graphs. Newman [2010] made a detailed introduction to these three centrality metrics as well as other node centrality measures, for example, Katz centrality, PageRank, Hubs, and Authorities. Other works on axiomatizing these centrality metrics for graphs were conducted by Garg [2009] and Boldi and Vigna [2014].

Many real-world graph applications and mining tasks leverage *network/graph similarity*, which is usually measured using three fundamental approaches in network science: *structural equivalence*, *automorphic equivalence*, and *regular equivalence* [Similarity 2015; Newman 2010]. Among the three approaches, structural equivalence is the strongest form of similarity. In many real applications, it is usually relaxed to some weak form of similarity [Similarity 2015; Hanneman and Riddle 2005]. Along this line and leveraging different structural characteristics of nodes within a graph, several *structural similarity* measures have been developed in various applications [Basak et al. 1988; Brandes and Lerner 2004; Zhou et al. 2009; Narayanan and Shmatikov 2009; Nilizadeh et al. 2014]. For instance, Basak et al. [1988] studied to determine the structural similarity of chemicals using graph-theoretic indices. Brandes and Lerner [2004] introduced structural similarity by relaxation of equitable partitions and applied the concept for role assignment (role extraction). Zhou et al. [2009] defined structural similarity based on neighborhood random walk distance and applied the concept for graph clustering. In the security and privacy area, structural similarity is also widely used to measure how similar two users are with respect to different graph theory metrics (e.g., degree centrality) [Narayanan and Shmatikov 2009; Nilizadeh et al. 2014].

2.2. Anonymize Social and Mobility Data

Social and mobility trace data are now easily obtainable and available through multiple channels, for example, academic and government data mining, advertising, third-party applications, data aggregation and inferring, and privacy attack and acquiring [Backstrom et al. 2007; Narayanan and Shmatikov 2009; Srivatsa and Hicks 2012]. To protect the privacy of publicly released data, a common method is to anonymize data by removing PII (e.g., name, age, social security number) before releasing data. However, this naive data anonymization is usually vulnerable to de-anonymization attacks [Campan and Truta 2008; Hay et al. 2008; Liu and Terzi 2008]. 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 nondistinguishable and thus defend against the de-anonymization (reidentification) attacks.

To preserve the privacy of sensitive relationships in graph data, Zheleva and Getoor [2007] designed five different privacy preservation strategies depending on the amount of data removed and the amount of privacy preserved. However, the common availability of auxiliary information for an adversary is not taken into account in the designed strategies.

Hay et al. [2008] introduced *k*-anonymity to social data anonymization. An assumption made on an 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 [Narayanan and Shmatikov 2009]. Nevertheless, the fact is, social graphs' average degree tends to be large and still increasing [McAuley and Leskovec 2012; Gong et al. 2012b]. 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.

Campan and Truta [2008] extended the k-anonymity scheme in Hay et al. [2008] 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 toward ID anonymization on graphs by Liu and Terzi [2008], where the a 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 Narayanan and Shmatikov [2009], the cornerstone of k-anonymity is based on data's syntactic property, which may not work on protecting actual data privacy even been satisfied.

2.3. De-Anonymize Social and Mobility Data

The most closely related works to this article are Backstrom et al. [2007], Narayanan and Shmatikov [2009], Srivatsa and Hicks [2012], and Ji et al. [2014]. Backstrom et al. [2007] introduced both active attacks and passive attacks to de-anonymize social data. For the active attack, the adversary should create a number of Sybil nodes and build relationships between Sybil nodes and target nodes before data release (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 [Narayanan and Shmatikov 2009], 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 data continues to increase [McAuley and Leskovec 2012; Gong et al. 2012b]. To execute active attack, many Sybil nodes and relationships/ties should be created, which is not practical. Furthermore, Sybil defense schemes [Alvisi et al. 2013; Yu et al. 2008a, 2008b, 2009] 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 Backstrom et al. [2007], adversaries can breach the privacy of users with whom they are linked, which is again suitable for small social networks and difficult to extend to large-scale social data.

Narayanan and Shmatikov [2009] extended the de-anonymization attack to large-scale directed social network data; that is, the social data carries direction information, which can be used as auxiliary knowledge. The designed de-anonymization algorithm included 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 Narayanan and Shmatikov [2009] is $O((|E_1| + |E_2|)d_1d_2) = O(n^4)$, where $|E_1|$ and d_1 ($|E_2|$ and d_2, respectively) are the edge set cardinality and degree bound of the anonymized graph (auxiliary graph, respectively), respectively, and n is the number of nodes in the anonymized graph or auxiliary graph (same from the order perspective).

Srivatsa and Hicks [2012] 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 subgraph 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 (e.g., to de-anonymize the Smallblue dataset, which has 125 nodes [Smallblue 2009] with five 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 Srivatsa and Hicks [2012].

Ji et al. [2014] 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. [2015a] further theoretically studied the de-anonymizability of social networks with seed knowledge. 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 [Ji et al. 2015b; SecGraph 2015]. In SecGraph, they implemented and evaluated 11 graph anonymization schemes, 12 graph utility metrics, seven application utility metrics, and 15 modern graph de-anonymization attacks (including the two attacks proposed in this article). 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. Nilizadeh et al. [2014] studied how to use the graph community information to enhance existing seed-based de-anonymization attacks (e.g., Narayanan and Shmatikov [2009] and Srivatsa and Hicks [2012]). They proposed a community-based de-anonymization framework, which de-anonymizes a graph first at the community level and then at the user level.

2.4. 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 that 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 there are 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 (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 Narayanan and Shmatikov [2009] applies to social 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 Narayanan and Shmatikov [2009]. 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.

3. PRELIMINARIES AND SYSTEM MODEL

3.1. Anonymized and Auxiliary Data Models

3.1.1. Anonymized Data Graph. In this article, we consider the anonymized data that can be modeled by an undirected graph,[2] 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), $E^a = \{l^a_{i,j} | i, j \in V^a, \text{ and there is a tie between } i \text{ and } j\}$ 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 the 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 $\mathbb{P}^a_{i,j} = \{p^a(i, j)\}$ as the set of all the shortest paths between i and j. Furthermore, we define the diameter of G^a as $D^a = \max\{|p^a(i, j)| \forall i, j \in V^a, p^a(i, j) \in \mathbb{P}^a_{i,j}\}$, that is, the length of the longest shortest path in G^a.

3.1.2. Auxiliary Data Graph. As in Narayanan and Shmatikov [2009] and Srivatsa and Hicks [2012], we assume the auxiliary data is the information crawled in current online social networks, for example, the "follow" relationships on Twitter [Narayanan and Shmatikov 2009], the "contact" relationships on Flickr, the "friend" relationships on Facebook, and the "circle" relationships on Google+. Furthermore, similar to 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 the 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 for 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 Section 5.

3.2. 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 accurately as possible. Then, adversaries can rely

[2] Note that the de-anonymization algorithm designed in this article 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 Narayanan and Shmatikov [2009].

Table I. 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

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 $\Phi : G^a \rightarrow G^u$, 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

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

be the outcome of a de-anonymization attack such that

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

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

3.3. Datasets

In this article, we employ six well-known datasets to examine the effectiveness of the designed de-anonymization framework: St Andrews/Facebook [Bigwood et al. 2011; Srivatsa and Hicks 2012], Infocom06/DBLP [Scott et al. 2009; Srivatsa and Hicks 2012], Smallbule/Facebook [Smallblue 2009; Srivatsa and Hicks 2012], ArnetMiner [Tang et al. 2008], Google+ [Gong et al. 2012b], and Facebook [Viswanath et al. 2009]. St Andrews, Infocom06, and Smallbule are three mobility trace datasets. The St Andrews dataset contains the 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 4 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 I. We employ the exact same techniques as in the previous work [Srivatsa and Hicks 2012] to preprocess the three mobility trace datasets to obtain three anonymized data graphs. To de-anonymize the aforementioned three anonymized mobility data traces, we employ three auxiliary social network datasets [Srivatsa and Hicks 2012] 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 datasets: ArnetMiner [Tang et al. 2008], Google+ [Gong et al. 2012a], and Facebook [Viswanath et al. 2009]. ArnetMiner is an online academic social network. In this article, the employed data is extracted from ArnetMiner in 2011 on the 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, 2011 [Gong et al. 2012a], 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 [Gong et al. 2012a], some connections that appeared in AUG may not appear in JUL and vice 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 section (Section 6).

4. 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 rationality and feasibility of seed selection in our de-anonymization framework (as well as other de-anonymization attacks) lie in three realities. The first is the common availability of huge amounts of social data, which is an open and rich source for obtaining a small number of seeds. For instance, (1) 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 [Narayanan and Shmatikov 2008]; (2) the social 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 [Narayanan and Shmatikov 2009]; and (3) online social network operators (e.g., Facebook, Twitter) and researchers (e.g., Stanford SNAP Datasets [SNAP 2014], Dartmouth CRAWDAD [CRAWDAD 2014]) publish many kinds of anonymized/unanoymized social data periodically. 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: (1) seed mapping information could be acquired due to data leakage—for example, some data may be leaked in data release for academic and government data mining with the original purpose of assisting research [Narayanan and Shmatikov 2008, 2009]; (2) auxiliary information can be collected by launching third-party applications on online social networks (many successful examples are surveyed in Narayanan and Shmatikov [2009]); (3) considering the common availability of huge amounts of social 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 [Srivatsa and Hicks 2012]; and (4) it is also possible to obtain a small number of seed mappings in a human-assisted semiautomatic manner. Adversaries can crawl some data first and then rely on human-assisted semiautomatic analysis to obtain some auxiliary information [Narayanan and Shmatikov 2008]. 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 [Backstrom et al. 2007; Narayanan and Shmatikov 2009; Srivatsa and Hicks 2012] 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* [Backstrom et al. 2007], 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 [Yu et al. 2008a, 2009], 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 [Backstrom et al. 2007; Narayanan and Shmatikov 2009; Srivatsa and Hicks 2012]. 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 [Singh et al. 2009; Hornyack et al. 2011; Egele et al. 2011]. (4) Some other existing attacks and seed identifying algorithms can be employed for seed selection, for example, the seed selection used in Backstrom et al. [2007] for active and passive attacks, and the clique-based seed identification in Narayanan and Shmatikov [2009].

Since seed selection is not our primary contribution in this article, we assume we have identified κ seed mappings by exploiting the aforementioned strategies individually or collaboratively, denoted by $\mathcal{M}_s = \{(s_1, s'_1), (s_2, s'_2), \ldots, (s_\kappa, s'_\kappa)\}$, 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 $\mathcal{M}_k\,(k = 1, 2, \ldots)$ be the mapping set after the kth iteration. To facilitate our discussion, we first define some terminologies as follows.

Let $M_k^a = \bigcup_{i=1}^{|\mathcal{M}_k|}\{v_i|(v_i, v_i') \in \mathcal{M}_k\}$ and $M_k^u = \bigcup_{i=1}^{|\mathcal{M}_k|}\{v_i'|(v_i, v_i') \in \mathcal{M}_k\} \setminus \{\perp\}$ 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 $\exists v_i \in M_k^a\, s.t.\, v_j \in N^a(v_i)\}$; that is, $\Lambda^1(M_k^a)$ denotes the set of nodes in G^a that have some neighbors that have been mapped and themselves have 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 $\exists v_i \in M_k^a\, s.t.\, |p^a(v_i, v_j)| \leq \delta\}$; that is, $\Lambda^\delta(M_k^a)$ denotes the set of nodes in G^a that are at most δ hops away from some node that has been mapped and that themselves have not been mapped yet. Here, $\delta(\delta = 1, 2, \ldots)$ is called the *spanning factor* in the mapping 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 that 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 properties,* and *inheritance properties* in the following subsections 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 [Freeeman 1978; Newman 2010; Similarity 2015]. In this article, 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* [Freeeman 1978; Newman 2010; Similarity 2015]. In the case that the considering data is modeled by a weighted graph, we employ the weighted versions [Opsahl et al. 2010; Newman 2010] of the employed three centrality measurements.

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, that is, the number of links with this node as an endpoint [Freeeman 1978; Newman 2010]. 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 $d_{v'} = \Delta_{v'}^u = |N^u(v')|$. To show some examples, 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 Figure 2. From Figure 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 Figure 1, which consists of six nodes and seven 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

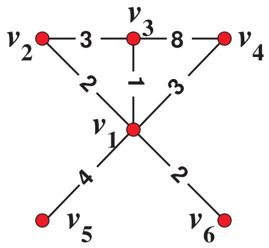
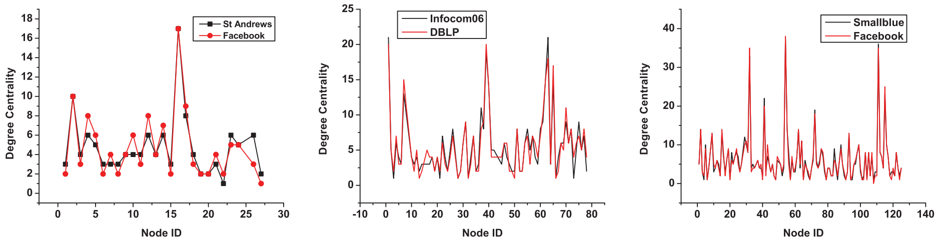


Fig. 1. A weighted graph.



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

Fig. 2. Degree centrality.

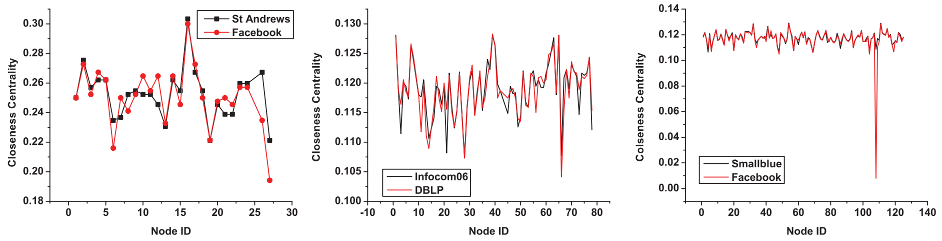
[Opsahl et al. 2010]. For instance, $d_{v_2} = d_{v_4}$ in Figure 1. However, the links associated with v_2 and v_4 have different weights or sum weights, which cannot be reflected by d_{v_2} and d_{v_4}. One naive idea is to define the degree centrality of a node in a weighted graph as the sum of the weights on the links associated with that node [Opsahl et al. 2010]. Nevertheless, this definition overlooks the information about the number of links associated with a node on the other hand. As shown in Figure 1, $\sum_{j \neq 1} w_{1,j} = \sum_{j \neq 3} w_{3,j} = 12$, while $d_{v_1} \neq d_{v_3}$ (as defined in Section 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 employ the *weighted degree centrality* definition proposed in Opsahl et al. [2010]. Formally, for $v \in V^a$, its weighted degree centrality is

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

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



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

Fig. 3. Closeness centrality.

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 [Freeeman 1978; Newman 2010; Similarity 2015]. 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

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

As an example, Figure 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 Figure 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*[3] for $v \in V^a$ and $v' \in V^u$ as

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

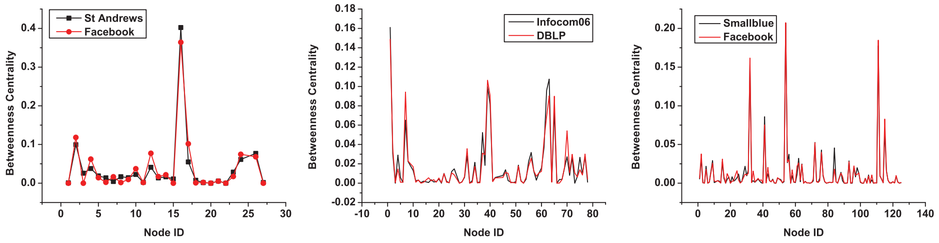
and

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

respectively, where $p_w^a(\cdot, \cdot)/p_w^u(\cdot, \cdot)$ 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 times a node acts as a bridge (intermediate node) along the shortest path between two other nodes [Freeeman 1978;

[3] Note that weighted closeness centrality is a concept in network science and there are other definitions for this term in Opsahl et al. [2010] and Newman [2010].



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

Fig. 4. Betweenness centrality.

Newman 2010; Similarity 2015]. Formally, for $v \in V^a$, its *betweenness centrality* b_v in G^a is defined as

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

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 $p^a(x, y)\}|$ that 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

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

According to the definition, we obtain the betweenness centrality of nodes in St Andrews/Facebook, Infocom06/DBLP, and Smallblue/Facebook as shown in Figure 4. From Figure 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*[4] for $v \in V^a$ and $v' \in V^u$ as

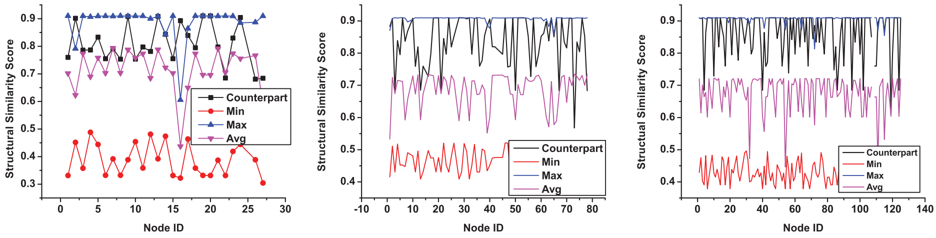
$$wb_v = \frac{\sum_{x \neq v \neq y} \frac{\sigma^{wa}_{xy}(v)}{\sigma^{wa}_{xy}}}{\binom{|V^a|-1}{2}} = \frac{2}{(|V^a|-1)(|V^a|-2)} \cdot \sum_{x \neq v \neq y} \frac{\sigma^{wa}_{xy}(v)}{\sigma^{wa}_{xy}} \tag{13}$$

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

[4] Similar to the weighted closeness centrality, weighted betweenness centrality is also a concept in network science and there are other definitions for this term in Opsahl et al. [2010] and Newman [2010].



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

Fig. 5. Structural similarity.

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,*[5] 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 $\mathbf{S}^u(v')$[6], that is,

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

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 Figure 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 $\frac{1}{|V^u|-1} \sum_{x' \in V^u, x' \neq \gamma(v)} s_S(v, x')$. From Figure 5, we have the following two basic observations:

—For some nodes with distinguished structural characteristics (e.g., nodes 2, 16, and 24 in St Andrews; nodes 10 and 40 in Infocom06; and nodes 19, 54, 64, 72, 111, and

[5] Depending on the specific applications, several other structural similarity measures were defined, for example, chemicals' structure comparison [Basak et al. 1988], role assignment/extraction [Brandes and Lerner 2004], and graph clustering [Zhou et al. 2009]. In this article, we define the structural similarity measure mainly for graph de-anonymization. All these measures can be considered as different relations of *structural equivalence*, an approach of defining *graph/network similarity* in network science [Newman 2010; Hanneman and Riddle 2005; Similarity 2015].

[6] Note that it is also possible to employ other metrics to measure the structural similarity between two nodes, for example, the Euclidian distance. Nevertheless, when different levels of noise are added to the anonymized graph (e.g., add/delete edges), the Euclidian distance between two vectors may vary significantly, while their distribution similarity is more stable (during the graph anonymization process). Thus, to maximize robustness, we employ cosine similarity.

115 in Smallblue), they agree with their counterparts and disagree with other nodes in the auxiliary graphs significantly (actually, they also show 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, and 26 in St Andrews; nodes 16, 73, and 78 in Infocom06; and nodes 4, 40, 86, 102, and 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 $\mathcal{M}_0 = \mathcal{M}_s = \{(s_1, s_1'), (s_2, s_2'), \ldots, (s_\kappa, s_\kappa')\}$. This a priori 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,*[7] denoted by $\mathbf{D}^a(v)$ to the seeds in $M_0^a = \{s_1, s_2, \ldots, s_\kappa\}$, as

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

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 $M_0^u = \{s_1', s_2', \ldots, s_\kappa'\}$ 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

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

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 $\mathbf{D}^u(v')$, that is,

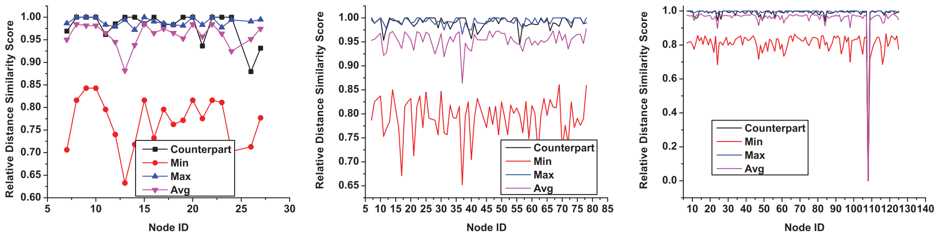
$$s_D(v, v') = \frac{\mathbf{D}^a(v) \cdot \mathbf{D}^u(v')}{\|\mathbf{D}^a(v)\| \|\mathbf{D}^u(v')\|}. \tag{20}$$

For St Andrews/Facebook, Infocom06/DBLP, and Smallblue/Facebook, by assuming $\mathcal{M}_s = \{(i, i)|i = 1, 2, \ldots, 6\}$ (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 the nodes in $V^u \setminus M_0^u$ as shown in Figure 6. From Figure 6, we can observe the following facts:

—Some anonymized nodes (which may be indistinctive with respect to structural similarity) (e.g., nodes 14, 19, and 23 in St Andrews; nodes 28, 31, 37, and 54 in Infocom06; and nodes 46, 63, 75, 98, and 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 high probability by employing the relative distance similarity based metric.

—For some nodes (e.g., nodes 11, 21, 26, and 27 in St Andrews; nodes 56 and 69 in Infocom06; and nodes 12, 13, and 26 in Smallblue), they are indistinctive on the relative distance similarity with respect to the initial seed selection $\{1, 2, 3, 4, 5, 6\}$.

[7] Note that the relative distance vector can also be defined using the Multidimensional Scaling (MDS) theory [Kruskal and Wish 1978]. To be consistent with existing anonymization/de-anonymization literature [Hay et al. 2008; Ji et al. 2014], 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

Fig. 6. Relative distance similarity.

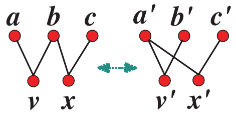


Fig. 7. Mapping inheritance.

To distinguish them, extra effort is expected, for example, 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.

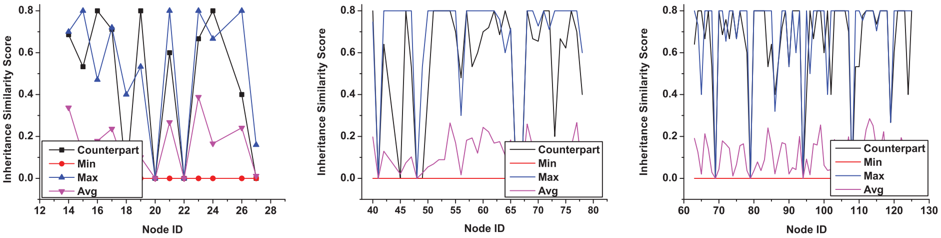
4.4. Inheritance Similarity

Besides the initial seed mapping, the de-anonymized nodes during each iteration, that is, $\mathcal{M}_k$, could provide further knowledge when de-anonymizing $\Lambda^{\delta}(M_k^a)$. As shown in Figure 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; that is, $\mathcal{M}_k$ 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

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

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 subsection.

From the definition of $s_I(v, v')$, we can see the following: (1) If two nodes have more common neighbors that have been mapped, then their inheritance similarity score is high. For example, in Figure 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. (2) 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. (3) We involve the similarity loss in counting $s_I(v, v')$, which



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

Fig. 8. Inheritance similarity.

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),[8] then the inheritance similarity between the rest of the nodes in the anonymized graph and the auxiliary graph is shown in Figure 8. From the result, we can observe that under the half number of nodes mapping assumption, some nodes (e.g., nodes 16, 19, and 24 in St Andrews; nodes 40, 56, 64, and 72 in Infocom06; and nodes 64, 65, 92, 96, 111, and 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 Figure 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 Figure 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 lack 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 US measurement by considering the structural similarity, relative

[8] Note that this assumption is used to illustrate an example. In a real attack, the inheritance similarity is calculated based on the nodes that have been de-anonymized. It could be any subset of the anonymized/auxiliary users.

distance similarity, and inheritance similarity synthetically for $v \in \Lambda^{\delta}(M_k^a)$ and $v' \in \Lambda^{\delta}(M_k^u)$ in the kth 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 DA framework, which is shown in Algorithm 1.

ALGORITHM 1: US Based **D**e-**A**nonymization (DA) Framework

```
   input: $G^a$, $G^u$, $\mathcal{M}_s$
   output: de-anonymization of $G^a$
1  $\mathcal{M}_0 = \mathcal{M}_s$, $k = 0$, $flag =$ true;
2  while $flag =$ true do
3      calculate $\Lambda^{\delta}(M_k^a)$ and $\Lambda^{\delta}(M_k^u)$;
4      if $\Lambda^{\delta}(M_k^a) = \emptyset$ or $\Lambda^{\delta}(M_k^u) = \emptyset$ then
5          output $\mathcal{M}_k$ and break;
6      for every $v \in \Lambda^{\delta}(M_k^a)$ do
7          for every $v' \in \Lambda^{\delta}(M_k^u)$ do
8              calculate $s(v, v')$;
9      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 $s(v, v')$;
10     use the Hungarian algorithm to obtain a maximum weighted bipartite matching of $B_k$,
       denoted by $\mathcal{M}' = \{(v, v') | v \in \Lambda^{\delta}(M_k^a), v' \in \Lambda^{\delta}(M_k^u)\}$;
11     for every $(x, x') \in \mathcal{M}'$ do
12         if $s(x, x') < \theta$ then
13             $\mathcal{M}' = \mathcal{M}' \setminus \{(x, x')\}$;
14     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
15         $\mathcal{M}' = \mathcal{M}' \setminus \{(x, x')\}$, i.e. onlykeep the Top-$K$ mapping pairs in $\mathcal{M}'$;
16     if $\mathcal{M}' = \emptyset$ then
17         output $\mathcal{M}_k$ and break;
18     $\mathcal{M}_{k+1} = \mathcal{M}_k \cup \mathcal{M}'$;
19     $k$++;
```

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 kth 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 the US score between the associated two nodes, that is, $w_{v,v'}^b = s(v, v')$. 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 $\Lambda^{\delta}(M_k^a)$ and nodes in $\Lambda^{\delta}(M_k^u)$ (lines 6–8). Subsequently, based on the obtained US scores, a weighted bipartite graph is constructed between nodes in $\Lambda^{\delta}(M_k^a)$ and nodes in $\Lambda^{\delta}(M_k^u)$ (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 $\mathcal{M}'$: (1) 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 importantly, 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). (2) We introduce a *mapping control factor* ϵ, or K equivalently, to limit the maximum number of mappings that have been recognized as correct de-anonymization (lines 14 and 15). During each iteration, only K mappings with the highest US scores will be taken as correct de-anonymization with confidence even if more mappings have 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 that have 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 and 19).

5. GENERALIZED SCALABLE DE-ANONYMIZATION

In this section, we extend DA to more general scenarios such as large-scale data de-anonymization, including the situation in which 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 is even if 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 ADA framework, denoted by ADA. ADA adaptively de-anonymizes G^a starting from a CMS, 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 $V_s^a = \bigcup_{x,y \in M_0^a} \{v | v \text{ lies on } p^a(x,y) \in \mathbb{P}^a(x,y)\}$ (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 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, that is, $G_c^a = G^a[V_c^a]$ (as shown in Figure 9). Similarly, we can define $V_s^u = \bigcup_{x',y' \in M_0^u} \{v' | v' \text{ lies on } p^u(x',y') \in \mathbb{P}^u(x',y')\}$ and $V_c^u = V_s^u \cup \Lambda^{\delta}(V_s^u)$. Then, the initial CMS on G^u is $G_c^u = G^u[V_c^u]$ (as shown in Figure 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

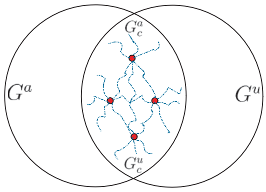


Fig. 9. Core matching subgraph (CMS). Initial seed mappings are denoted by red nodes.

Figure 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 **D**e-**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 $G_c^u$;
```

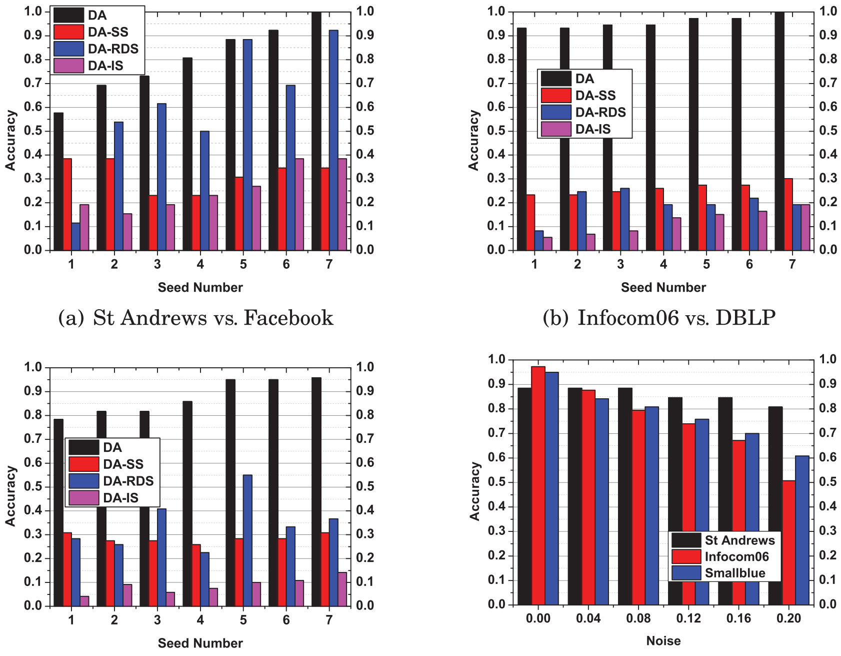
Now, based on CMS, we discuss ADA as shown in Algorithm 2. In Algorithm 2, μ is the *adaptive factor* that controls the spanning size of the CMS during each adaptive iteration. The basic idea of ADA is as follows. We start the de-anonymization from CMSs G_c^a and G_c^u by running DA (lines 1 and 2). 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 $\bigcup_i G_i^a$ and $\bigcup_j G_j^u$, 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.

6. EXPERIMENTS

In this section, 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 (ArnetMiner), and social data (Google+, Facebook).



(a) St Andrews vs. Facebook

(b) Infocom06 vs. DBLP

(c) Smallblue vs. Facebook

(d) De-anonymization with noise

Fig. 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.

6.1. De-Anonymize 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 Figures 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 Figures 10(a)–(c), we can see the following: (1) 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 seven seed mappings. (2) 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 distinguish a node from multiple perspectives, which is more efficient and comprehensive. As the analysis shows in Section 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 Figure 10(d) (on the knowledge of 5 seed mappings). In

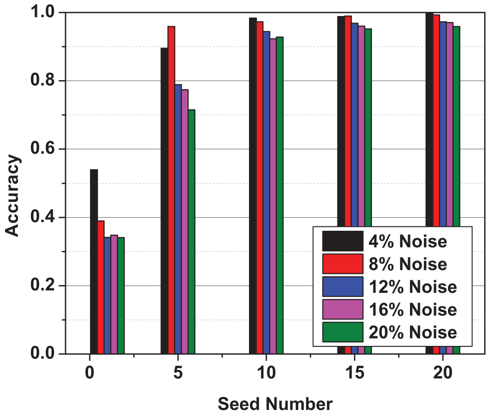


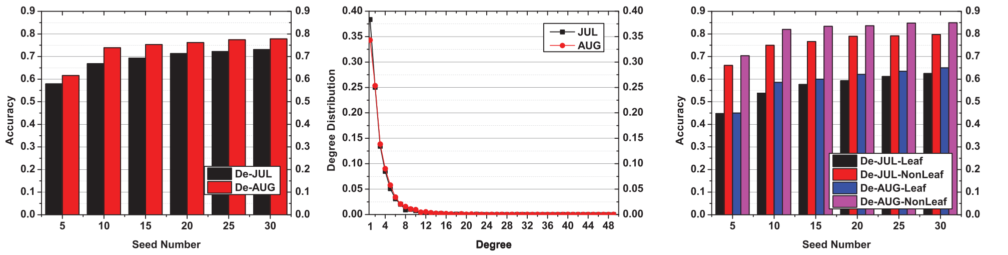
Fig. 11. De-anonymize ArnetMiner. Default parameter settings: $\alpha = 1.5$, $C = 0.9$, $c_S = 0.2$, $c_D = 0.6$, $c_I = 0.2$, $\theta = 0.6$, $\delta \in \{1, 2\}$, $\mu \in \{1, 2, 3\}$, and $\epsilon = 0.5$.

the experiment, we only add noise to the anonymized data. According to the same argument in Narayanan and Shmatikov [2009], 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 Figure 10(d), 20% of noise implies we add 10% spurious connections and delete 10% existing connections of $|E^a|$ from the anonymized data. From Figure 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 practice, 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 Figure 10(d), which confirms the assertion in Narayanan and Shmatikov [2009] that structure change may not provide effective privacy protection.

6.2. De-Anonymize 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 anonymize ArnetMiner by adding p percent noise as explained in the previous subsection. 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 Figure 11.

From Figure 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 one-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 five seed mappings are available and 92.8% accuracy if 10 seed



(a) De-anonymize Google+ (b) Degree distribution of Google+ (c) Degree distribution of Google+

Fig. 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 $\epsilon = 0.5$.

mappings are available. The presented de-anonymization framework is robust to noise on weighted data. When we have 10 or more seed mappings, the accuracy degradation of our de-anonymization algorithm is small even with more noise; for example, the accuracy is degraded from 99.7% in the 4% noise case to 96% in the 20% noise case. 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 dataset, 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, five to 10 seed mappings are sufficient.

6.3. De-Anonymize 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 Figure 12(a). Again, from Figure 12(a), we can see that the presented de-anonymization framework is very effective. Just based on the knowledge of five seed mappings, 57.9% of the users in JUL and 61.6% of the users in AUG can be successfully de-anonymized. 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 Figure 12(a): (1) 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 though the de-anonymization accuracy is around 70% for JUL and 75% for AUG; and (2) 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 that 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 Figure 12(b). From Figure 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 1, 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

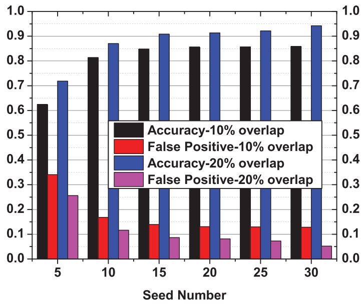


Fig. 13. De-anonymize Facebook. Default parameter settings: $C = 0.9$, $c_S = 0.2$, $c_D = 0.6$, $c_I = 0.2$, $\theta = 0.8$, $\delta \in \{1, 2\}$, $\mu \in \{1, 2, 3\}$, and $\epsilon = 0.5$.

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 if $x' = \gamma(x)$, $y' = \gamma(y)$, 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 at the experimental results. 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 1 or *nonleaf user set* if its degree is greater than 1. Then, we recalculate the de-anonymization accuracy for leaf users and nonleaf users, and the results are shown in Figure 12(c), where De-JUL-Leaf/De-AUG-Leaf represents the ratio of leaf nodes that have been successfully de-anonymized in JUL/AUG and De-JUL-NonLeaf/De-AUG-NonLeaf represents the ratio of nonleaf users that have been successfully de-anonymized in JUL/AUG. From Figure 12(c), we can see that (1) the successful de-anonymization ratio on nonleaf users is higher than that on leaf users in JUL and AUG—this is because nonleaf users carry more structural information; and (2) considering the results shown in Figure 12(a), the de-anonymization accuracy on nonleaf 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 Figure 12(c) confirm our argument that leaf users are more difficult than nonleaf users to de-anonymize. Furthermore, this is also why De-AUG has higher accuracy than De-JUL in Figure 12(a). AUG is easier to de-anonymize since it has fewer leaf users than JUL.

6.4. De-Anonymize 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 Narayanan and Shmatikov [2009], 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 Figure 13. As a fact, most of the existing de-anonymization attacks are not very effective for the scenario in which the overlap between the anonymized data and the auxiliary data is small or even cannot work totally. Surprisingly, for ADA, we can observe from Figure 13 that (1) based on the proposed CMS, ADA can successfully de-anonymize 62.4% of the common users with a false-positive error of 34.1% when the overlap is 10% and 71.8% of the common users with a false-positive error of 25.6% when the overlap is 20% with the knowledge of just five seed mappings; (2) the de-anonymization accuracy is improved to 81.3% (85.6%, respectively) and the false-positive error is decreased to 16.8% (13%, respectively) when the overlap is 10% and there are 10 (20, respectively) seed mappings available, and the de-anonymization accuracy is improved to 87% (90.8%, respectively) and the false-positive error is decreased to 11.6% (8.6%, respectively) when the overlap is 20% and there 10 (20, respectively) seed mappings available, which demonstrate that ADA is very effective in dealing with the partial data overlap situation; and (3) 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 Figure 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 practice.

7. CONCLUSION

In this article, 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 that take into account both local and global structural characteristics of data, the information obtained from auxiliary data, and the knowledge inherited from ongoing 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 data can be de-anonymized with coarse granularity. Subsequently, we introduce a Unified Similarity (US) measurement that 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 datasets: 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

L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi. 2013. SoK: The evolution of Sybil defense via social networks. In *S&P*.

L. Backstrom, C. Dwork, and J. Kleinberg. 2007. Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. In *WWW*.

S. C. Basak, V. R. Magnuson, G. J. Niemi, and R. R. Regal. 1988. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* 19 (1988), 17–44.

A. Bavelas. 1950. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am* 22 (1950), 725–730.

G. Bigwood, D. Rehunathan, M. Bateman, T. Henderson, and S. Bhatti. 2011. CRAWDAD dataset st_andrews/sassy. Retrieved from http:/crawdad.cs.dartmouth.edu/st_andrews/sassy.

P. Boldi and S. Vigna. 2014. Axioms for centrality. *Internet Math.* 10, 3–4 (2014), 222–262.

U. Brandes and J. Lerner. 2004. Structural similarity in graphs: A relaxation approach for role assignment. *LNCS* 3341 (2004), 184–195.

A. Campan and T. M. Truta. 2008. A clustering approach for data and structural anonymity in social networks. In *PinKDD*.

Centrality. 2015. Centrality - Wikipedia, the free encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Centrality.

CRAWDAD. 2014. Retrieved from http://crawdad.cs.dartmouth.edu/.

M. Egele, C. Kruegel, E. Kirda, and G. Vigna. 2011. PiOS: Detecting privacy leaks in iOS applications. *NDSS* (2011).

L. C. Freeeman. 1978. Centrality in social networks: Conceptual clarification. *Social Netw.* 1 (1978), 215–239.

M. Garg. 2009. Axiomatic foundations of centrality in networks. *Social Sci. Res. Netw.* (2009), 1–37.

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

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

R. Hanneman and M. Riddle. 2005. Introduction to Social Network Methods: Table of Contents. Retrieved from http://faculty.ucr.edu/∼hanneman/nettext/.

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

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

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

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

S. Ji, W. Li, P. Mittal, X. Hu, and R. Beyah. 2015b. SecGraph: A uniform and open-source evaluation system for graph data anonymization and de-anonymization. In *USENIX Security*.

J. Kruskal and M. Wish. 1978. *Multidimensional Scaling*. Sage Publications.

K. Liu and E. Terzi. 2008. Towards identity anonymization on graphs. *SIGMOD* (2008).

J. McAuley and J. Leskovec. 2012. Learning to discover social circles in ego networks. In *NIPS*.

A. Narayanan and V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets (de-anonymizing the Netflix prize dataset). In *S&P*.

A. Narayanan and V. Shmatikov. 2009. De-anonymizing social networks. In *S&P*.

M. E. J. Newman. 2010. *Networks: An Introduction.* Oxford University Press.

S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn. 2014. Community-enhanced de-anonymization of online social networks. In *CCS*.

T. Opsahl. 2010. Closeness centrality in networks with disconnected components | Tore Opsahl. Retrieved from http://toreopsahl.com/2010/03/20/closeness-centrality-in-networks-with-disconnected-components/.

T. Opsahl, F. Agneessens, and J. Skvoretz. 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Netw.* 32 (2010), 245–251.

Y. Rochat. 2009. Closeness centrality extended to unconnected graphs: The harmonic centrality index. In *ASNA*. 1–14.

J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. 2009. CRAWDAD dataset cambridge/haggle. Retrieved from http://crawdad.cs.dartmouth.edu/cambridge/haggle.

SecGraph. 2015. Retrieved from http://www.ece.gatech.edu/cap/secgraph/.

Similarity. 2015. Similarity (network science) - Wikipedia, the free encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Similarity_(network_science).

K. Singh, S. Bhola, and W. Lee. 2009. xBook: Redesigning privacy control in social networking platforms. In *USENIX*.

Smallblue. 2009. SmallBlue Research Projects. Retrieved from http://domino.research.ibm.com/comm/research_projects.nsf/pages/smallblue.index.html.

SNAP. 2014. Stanford Large Network Dataset Collection. Retrieved from http://snap.stanford.edu/data/.

M. Srivatsa and M. Hicks. 2012. Deanonymizing mobility traces: Using social networks as a side-channel. In *CCS*.

J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. 2008. ArnetMiner: Extraction and mining of academic social networks. In *KDD*.

B. Viswanath, A. N. Mislove, M. Cha, and K. P. Gummadi. 2009. On the evolution of user interaction in facebook. In *WOSN*.

H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. 2008a. SybilLimit: A near-optimal social network defense against Sybil attacks. In *S&P*.

H. Yu, M. Kaminsky, P. B. Gibbons, and A. D. Flaxman. 2008b. SybilGuard: Defending against Sybil attacks via social networks. *IEEE/ACM Trans. Netw.* 16, 3 (2008), 576–589.

H. Yu, C. Shi, M. Kaminsky, P. B. Gibbons, and F. Xiao. 2009. DSybil: Optimal Sybil-resistance for recommendation systems. In *S&P*.

E. Zheleva and L. Getoor. 2007. Preserving the privacy of sensitive relationships in graph data. In *PinKDD*.

Y. Zhou, H. Cheng, and J. X. Yu. 2009. Graph clustering based on structural/attribute similarities. In *VLDB*.

Received January 2015; revised November 2015; accepted February 2016