Structure based Data De-anonymization of Social Networks and Mobility Traces

Shouling Ji[1], Weiqing Li[1], Mudhakar Srivatsa[2], Jing S. He[3], and Raheem Beyah[1]

Georgia Institute of Technology[1], IBM T. J. Watson Research Center[2], KSU[3]
{sji, wli64}@gatech.edu, msrivats@us.ibm.com, jhe4@kennesaw.edu, rbeyah@ece.gatech.edu

Abstract. We present a novel *de-anonymization attack* on mobility trace data and social data. First, we design an Unified Similarity (US) measurement, based on which we present a US based De-Anonymization (DA) framework which iteratively de-anonymizes data with an accuracy guarantee. Then, to de-anonymize data *without the knowledge of the overlap size* between the anonymized data and the auxiliary data, we generalize DA to an Adaptive De-Anonymization (ADA) framework. Finally, we examine DA/ADA on mobility traces and social data sets.

1 Introduction

Social networking is a fast-growing business sector nowadays. To protect users' privacy, social network owners usually anonymize data by removing "Personally Identifiable Information (PII)" before releasing the data to the public. However, this data anonymization is vulnerable to a new *social auxiliary information based data de-anonymization attack* [1][2][3]. For example, just recently, it was reported that poorly anonymized logs revealed New York City cab drivers' detailed whereabouts (June 23, 2014) [5].

A few de-anonymization attacks have been designed for social data [1][2] or mobility trace data [3]. In [1], Backstrom et al. proposed active and passive attacks on social data. Since the proposed active attack relies on sybil nodes to obtain auxiliary information before social data release, it is not practical as analyzed in [2]. For the passive attack designed in [1], it is workable however not scalable [2]. In [2], Narayanan and Shmatikov showed a de-anonymization attack on social network data which can be modeled by directed graphs (where direction information carried by data can be viewed as free auxiliary information for adversaries). In [3], Srivatsa and Hicks presented the first de-anonymization attack to mobility traces by using social networks as a side-channel. Our work improves existing works in some or all of the following aspects. First, we significantly improve the de-anonymizaiton accuracy and decrease the computational complexity by proposing a novel *Core Matching Subgraphs* (CMS) based adaptive de-anonymization strategy. Second, besides utilizing nodes' local property, we incorporate nodes' 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 data sets. We give the detailed analysis and remarks in the related work (due to space limitation, we put the related work in the Technical Report [17] of this paper).

In summary, our main contributions are as follows[1]. (*i*) We define three de-anonymization metrics, namely *structural similarity*, *relative distance similarity*, and *inheritance similarity*, (*ii*) Toward effective de-anonymization, we define a *Unified Similarity* (US) measurement by collectively considering the defined structural similarity, relative distance similarity, and inheritance similarity. Subsequently, we propose a US based **D**e-**A**nonymization (DA) framework, by which we iteratively de-anonymize the anonymized data with an accuracy guarantee provided by a *de-anonymization threshold* and a *mapping control factor*. (*iii*) To de-anonymize data without the knowledge on the overlap size between the anonymized data and the auxiliary data, we generalize DA to an **A**daptive **D**e-**A**nonymization (ADA) framework. (*iv*) We apply the presented de-anonymization framework to mobility traces and social data sets. The experimental results demonstrate that the presented de-anonymization attack is very effective and robust. For instance, 93.2% of the users in Infocom06 [13] can be successfully de-anonymized given one seed mapping, and 58% of the users in Google+ can be de-anonymized given five seed mappings.

2 Preliminaries and Model

Anonymized Data Graph. In this paper, we consider anonymized data which 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 [10]), $E^a = \{l^a_{i,j} | i, j \in V^a$, and there is a tie between i 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+ [10]), and $W^a = \{w^a_{i,j} | i, j \in V^a, l^a_{i,j} \in E^a, w^a_{i,j}$ is a real number$\}$ is the set of possible weights associated with links in E^a (e.g., in a coauthor graph, the weight of a coauthor relationship could be the number of coauthored papers). If G^a is an unweighted graph, we simply define $w^a_{i,j} = 1$ for each link $l^a_{i,j} \in E^a$.

For $\forall i \in V^a$, we define its neighbor set as $N^a(i) = \{j \in V^a | l^a_{ij} \in E^a\}$. Then, $\Delta^a_i = |N^a(i)|$ represents the number of neighbors of i in G^a. For $\forall i, j \in V^a$, let $p^a(i, j)$ be a shortest path from i to j in G^a and $|p^a(i, j)|$ be the number of links on

[1] Due to the space limitation, we put more discussion and experimental results in the Technical Report [17] of this paper.

[2] Note that, the de-anonymization algorithm designed in this paper 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 [2] which could obtain better accuracy.

$p^a(i,j)$ (the number of links passed from i to j through $p^a(i,j)$). Then, we define $\mathbb{P}^a_{i,j} = \{p^a(i,j)\}$ the set of all the shortest pathes between i and j. Furthermore, we define the diameter of G^a as $D^a = \max\{|p^a(i,j)|\forall i,j \in V^a, p^a(i,j) \in \mathbb{P}^a_{i,j}\}$, i.e., the length of the longest shortest path in G^a.

Auxiliary Data Graph. As in [2][3][4], we assume the auxiliary data is the information crawled in current online social networks, e.g., the "follow" relationships on Twitter [2], the "friend" relationships on Facebook [3], etc. Furthermore, similar as the anonymized data, the auxiliary data can also be modeled as an undirected graph $G^u = (V^u, E^u, W^u)$, where V^u is the node set, E^u is set of all the links (relationships) among the nodes in V^u, and W^u is the set of possible weights associated with the links in E^u. As the definitions on the anonymized graph G^a, we can define the neighborhood of $\forall i \in V^u$ as $N^u(i)$, the shortest path set between $i \in V^u$ and $j \in V^u$ as $\mathbb{P}^u(i,j) = \{p^u(i,j)\}$, and the diameter of G^u as $D^u = \max\{|p^u(i,j)|\forall i,j \in V^u, p^u(i,j) \in \mathbb{P}^u(i,j)\}$.

In addition, we assume G^a and G^u are connected. Note that this is not a limitation of our scheme. The designed de-anonymization attack is also applicable to the case where G^a or G^u is not connected. We will discuss this in Section 4.

Attack Model. Our de-anonymization objective is to map the nodes in the anonymized graph G^a to the nodes in the auxiliary graph G^u as accurate as possible. Formally, let $\gamma(v)$ be the *objective reality* of $v \in G^a$ in the physical world. Then, an ideal de-anonymization can be represented by mapping $\Phi : G^a \to G^u$, such that for $v \in G^a$, $\Phi(v) = v'$ if $v' = \Phi(v) \in V^u$ and $\Phi(v) = \perp$ if $\Phi(v) \notin V^u$, where $\perp$ is a special *not existing indicator* in the auxiliary data graph. Now, let $\mathcal{M} = \{(v_1, v'_1), (v_2, v'_2), \cdots, (v_n, v'_n)\}$ be the outcome of a de-anonymization attack such that $v_i \in V^a, \cup v_i = V^a, n = |V^a|$ $(i = 1, 2, \cdots, n)$ and $v'_i = \Phi(v_i), v'_i \in V^u \cup \{\perp\}$ $(i = 1, 2, \cdots, n)$. Then, the de-anonymization on v_i is said to be *successful* if $\Phi(v_i) = \gamma(v_i)$ when $\gamma(v_i) \in V^u$ or $\Phi(v_i) = \perp$ when $\gamma(v_i) \notin V^u$; and *failure* if $\Phi(v_i) \in \{u|u \in V^u, u \neq \gamma(v_i)\} \cup \{\perp\}$ when $\gamma(v_i) \in V^u$ or $\Phi(v_i) \neq \perp$ when $\gamma(v_i) \notin V^u$. In this paper, we are aiming to design a de-anonymization attack with a high success rate (accuracy).

3 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.

Seed Selection and Mapping Spanning. Seed selection is possible in our de-anonymization framework because of three reasons. The first reason 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, the data published for academic and government data mining may also release some auxiliary information [4]. The second reason is the existence of multiple effective channels to obtain

a small number of seed mappings (actually, we can obtain much richer auxiliary information), e.g., data leakage [2][4], third party applications [2], etc. The third reason 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 [1][2][3][4] individually or collaboratively, e.g., launching a small scale *sybil attack* [1][2], compromising a small number of nodes [1][2][3], third party applications [6][7][8], etc.

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

Let $M_k^a = \bigcup_{i=1}^{|\mathcal{M}_k|} \{v_i | (v_i, v'_i) \in \mathcal{M}_k\}$ 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 until iteration k in G^a and G^u, respectively. Then, we define the *1-hop mapping spanning set* of M_k^a as $\Lambda^1(M_k^a) = \{v_j \in V^a | v_j \notin M_k^a \text{ and } \exists v_i \in M_k^a \ s.t. \ v_j \in N^a(v_i)\}$, i.e., $\Lambda^1(M_k^a)$ denotes the set of nodes in G^a that have some neighbor been mapped and themselves not been mapped yet. To be general, we can also define the *δ-hop mapping spanning set* of M_k^a as $\Lambda^\delta(M_k^a) = \{v_j \in V^a | v_j \notin M_k^a \text{ and } \exists v_i \in M_k^a \ s.t. \ |p^a(v_i, v_j)| \leq \delta\}$, i.e., $\Lambda^\delta(M_k^a)$ denotes the set of nodes in G^a that are at most δ hops away from some node been mapped and themselves not been mapped yet. Here, $\delta (\delta = 1, 2, \cdots)$ is called the *spanning factor* in the mapping 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 $\Lambda^1(M_k^u) = \{v'_j \in V^u | v'_j \notin M_k^u \text{ and } \exists v'_i \in M_k^u \ s.t. \ v'_j \in N^u(v'_i)\}$ and $\Lambda^\delta(M_k^u) = \{v'_j \in V^u | v'_j \notin M_k^u \text{ and } \exists v'_i \in M_k^u \ s.t. \ |p^u(v'_i, v'_j)| \leq \delta\}$, respectively. Based on the defined δ-hop mapping sets $\Lambda^\delta(M_k^a)$ and $\Lambda^\delta(M_k^u)$, we try to seek a mapping Φ which maps the anonymized nodes in $\Lambda^\delta(M_k^a)$ to some nodes in $\Lambda^\delta(M_k^u) \cup \{\perp\}$ iteratively in the mapping propagation phase of our de-anonymization framework.

Structural Similarity. Since both anonymized data and the auxiliary data can be modeled by graphs, the structural/topological characteristics could be a reference for *coarse granularity* (*high level*) de-anonymization. Here, coarse granularity de-anonymization implies for an anonymized node $v \in V^a$, we de-anonymize it by mapping it to some nodes $\{v' | v' \in V^u \cup \{\perp\}$ and v' in G^u is structurally similar to v in $G^a\}$ even the ideal *one-to-one mapping* cannot be achieved. Structural characteristics based coarse granularity de-anonymization

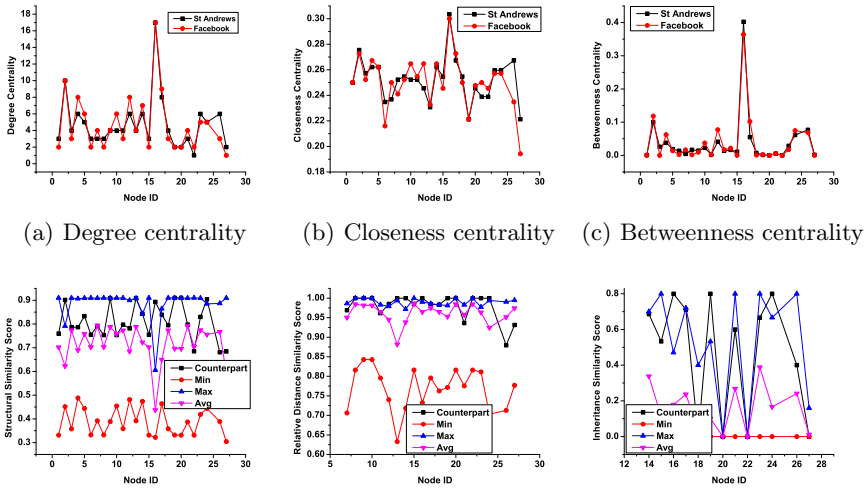
is meaningful since we can employ further techniques to refine the coarse granularity de-anonymization and finally de-anonymize v exactly.

In graph theory, the concept of *centrality* to measure the topological importance and characteristic of a node within a graph is often used. In this paper, we employ three centrality measurements to capture the topological property of a node in G^a or G^u, namely *degree centrality*, *closeness centrality*, and *betweenness centrality*. In the case that the considered data is modeled by a weighted graph, we also define the weighted version of the three centrality measurements. Furthermore, to demonstrate the aforementioned topological properties, we will employ an example data set (St Andrews [11]), which consists of a mobility trace data set of 27 users and a Facebook network of the same 27 users. The mobility trace data set has 18,241 WiFi records. We use the method in [3] to construct an anonymized graph on the 27 users based on the mobility trace data and take the Facebook network as the auxiliary data.

Degree Centrality and Weighted Degree Centrality. The *degree centrality* is defined as the number of ties that a node has in a graph. For instance, in the considered anonymized data graph, the degree centrality of $v \in V^a$ is defined as $d_v = \Delta_v^a = |N^a(v)|$. We calculate the degree centrality of the nodes in St Andrews and their counterparts in Facebook, and the result is shown in Fig.1 (a). From Fig.1 (a), we observe that the degree centrality distributions of the anonymized graph and auxiliary graph are similar, which implies 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 can be used for coarse granularity de-anonymization.

When the data being considered is modeled by a weighted graph, the weights on links provide extra information in characterizing the centrality of a node. In this case, the degree centrality defined for an unweighted graph cannot properly reflect a node's structural importance [9]. To consider both the number of links associated with a node and the weights on these links, we define the *weighted degree centrality* for $v \in V^a$ as $wd_v = \Delta_v^a(\frac{\sum_{u \in N^a(v)} w_{v,u}^a}{\Delta_v^a})^\alpha$, where α is a positive tuning parameter that can be set according to the research setting and data [9]. Basically, when $0 \leq \alpha \leq 1$, high degree is considered more important, whereas when $\alpha \geq 1$, weight is considered more important. Similarly, we can define the *weighted degree centrality* for $v' \in V^u$ as $wd_{v'} = \Delta_{v'}^u(\frac{\sum_{u' \in N^u(v')} w_{v',u'}^u}{\Delta_{v'}^u})^\alpha$.

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



(a) Degree centrality (b) Closeness centrality (c) Betweenness centrality

(d) $s_S(v, v')$ (e) $s_D(v, v')$ (f) $s_I(v, v')$

Fig. 1. Structural/relative distance/inheritance similarity.

$v \in V^a$, its *closeness centrality* c_v is defined as $c_v = \frac{|V^a|-1}{\sum_{u \in V^a, u \neq v} |p^a(v,u)|}$. Similarly, the *closeness centrality* $c_{v'}$ of $v' \in V^u$ is defined as $c_{v'} = \frac{|V^u|-1}{\sum_{u' \in V^u, u' \neq v'} |p^u(v',u')|}$.

Fig.1 (b) demonstrates the closeness centrality score of the nodes in St Andrews and their counterparts in the corresponding social graph Facebook. From Fig.1 (b), 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 generalize the *weighted closeness centrality* for $v \in V^a$ and $v' \in V^u$ as $wc_v = \frac{|V^a|-1}{\sum_{u \in V^a, u \neq v} |p^a_w(v,u)|}$ and $wc_{v'} = \frac{|V^u|-1}{\sum_{u' \in V^u, u' \neq v'} |p^u_w(v',u')|}$, respectively, where $p^a_w(\cdot,\cdot)/p^u_w(\cdot,\cdot)$ is the shortest path between two nodes in a weighted graph.

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. Formally, for $v \in V^a$, its *betweenness centrality* b_v in G^a is defined as $b_v = \frac{2}{(|V^a|-1)(|V^a|-2)} \cdot \sum_{x \neq v \neq y} \frac{\sigma^a_{xy}(v)}{\sigma^a_{xy}}$, where $x', y' \in V^a$, $\sigma^a_{xy} = |\mathbb{P}^a(x,y)|$ is the number of all the shortest paths between x and y in G^a, and $\sigma^a_{xy}(v) = |\{p^a(x,y) \in \mathbb{P}^a(x,y) | v$ is an intermediate node on path $p^a(x,y)\}|$ is the number of shortest paths between x and y in G^a that v lies on. Similarly, the *betweenness centrality* $b_{v'}$ of $v' \in V^u$ in G^u is defined as $b_{v'} = \frac{2}{(|V^u|-1)(|V^u|-2)} \cdot \sum_{x' \neq v' \neq y'} \frac{\sigma^u_{x'y'}(v')}{\sigma^u_{x'y'}}$.

According to the definition, we obtain the betweenness centrality of nodes in St Andrews as shown in Fig.1 (c). From Fig.1 (c), the nodes in G^a and their counterparts in G^u agree highly on betweenness centrality. Consequently, betweenness centrality can also be employed in our de-anonymization framework for distinguishing mappings. For the case that the considering data is modeled as a weighted graph, we define the *weighted betweenness centrality* for $v \in V^a$ and $v' \in V^u$ as $wb_v = \frac{2}{(|V^a|-1)(|V^a|-2)} \cdot \sum_{x \neq v \neq y} \frac{\sigma_{xy}^{wa}(v)}{\sigma_{xy}^{wa}}$ and $wb_{v'} = \frac{2}{(|V^u|-1)(|V^u|-2)} \cdot \sum_{x' \neq v' \neq y'} \frac{\sigma_{x'y'}^{wu}(v')}{\sigma_{x'y'}^{wu}}$, respectively, where σ_{xy}^{wa} and $\sigma_{xy}^{wa(v)}$ (respectively, $\sigma_{x'y'}^{wu}$ and $\sigma_{x'y'}^{wa(v')}$) are the number of shortest paths between x and y (respectively, x' and y') and the number of shortest paths between x and y (respectively, x' and y') passing v (respectively, v') in the weighted graph G^a (respectively, G^u).

Structural Similarity. From the analysis on real data sets, the local and global structural characteristics carried by degree, closeness, and betweenness centralities of nodes can guide our de-anonymization framework design. Following this direction, to consider and utilize nodes' structural property integrally, we define a unified structural measurement, namely *structural similarity*, to jointly count two nodes' both local and global topological properties. First, for $v \in V^a$ and $v' \in V^u$, we define two *structural characteristic vectors* $\mathbf{S}^a(v)$ and $\mathbf{S}^u(v')$ respectively in terms of their (weighted) degree, closeness, and betweenness centralities as follows: $\mathbf{S}^a(v) = [d_v, c_v, b_v, wd_v, wc_v, wb_v]$ and $\mathbf{S}^u(v') = [d_{v'}, c_{v'}, b_{v'}, wd_{v'}, wc_{v'}, wb_{v'}]$. In $\mathbf{S}^a(v)$, if G^a is unweighted, we set $wd_v = wc_v = wb_v = 0$; otherwise, we first count d_v, c_v, and b_v by assuming G^a is unweighted, and then count wd_v, wc_v, and wb_v in the weighted G^a. We also apply the same method to obtain $\mathbf{S}^u(v')$ in G^u. Based on $\mathbf{S}^a(v)$ and $\mathbf{S}^u(v')$, we define the *structural similarity* between $v \in V^a$ and $v' \in V^u$, denoted by $s_S(v, v')$, as the *cosine similarity* between $\mathbf{S}^a(v)$ and $\mathbf{S}^u(v')$, i.e., $s_S(v, v') = \frac{\mathbf{S}^a(v) \cdot \mathbf{S}^u(v')}{\|\mathbf{S}^a(v)\| \|\mathbf{S}^u(v')\|}$, where $\cdot$ is the *dot product* and $\|\cdot\|$ is the *magnitude* of a vector.

The structural similarity between the nodes in St Andrews and its auxiliary network Facebook is shown in Fig.1 (d), 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 Fig.1 (d), we have the following two basic observations. (*i*) For some nodes with distinguished structural characteristics, e.g., nodes 2, 16, 24, they agree with their counterparts and disagree with other nodes in the auxiliary graphs significantly. 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. (*ii*) For the nodes with indistinctive structural similarities, e.g., nodes 7, 10, 22, 26, exact node mapping relying on structural property alone is difficult or impossible to achieve from the view of graph the-

ory. Fortunately, even if this is true, structural characteristics can also help us to differentiate these indistinctive nodes from most of the other nodes. Hence, structural similarity based coarse granularity de-anonymization is practical.

Relative Distance Similarity. In the first phase, we select an initial seed mapping $\mathcal{M}_0 = \mathcal{M}_s = \{(s_1, s'_1), (s_2, s'_2), \cdots, (s_\kappa, s'_\kappa)\}$. This apriori knowledge can be used to conduct more confident ratiocination in de-anonymization. Therefore, for $v \in V^a \setminus M_0^a$, we define its *relative distance vector*, denoted by $\mathbf{D}^a(v)$ to the seeds in $M_0^a = \{s_1, s_2, \cdots, s_\kappa\}$ as $\mathbf{D}^a(v) = [D_1^a(v), D_2^a(v), \cdots, D_\kappa^a(v)]$, where $D_i^a(v) = \frac{|p^a(v,s_i)|}{D^a}$ is the *normalized relative distance* between v and seed s_i. Similarly, based on the initial seed set $M_0^u = \{s'_1, s'_2, \cdots, s'_\kappa\}$ in G^u, we can define the *relative distance vector* for $v' \in V^u \setminus M_0^u$ to the seeds in M_0^u as $\mathbf{D}^u(v') = [D_1^u(v'), D_2^u(v'), \cdots, D_\kappa^u(v')]$, where $D_i^u(v') = \frac{|p^u(v',s'_i)|}{D^u}$ is the *normalized relative distance* between v' and seed s'_i. Again, we can define the *relative distance similarity* between $v \in V^a \setminus M_0^a$ and $v' \in V^u \setminus M_0^u$, denoted by $s_D(v, v')$, as the *cosine similarity* between $\mathbf{D}^a(v)$ and $\mathbf{D}^u(v')$, i.e., $s_D(v, v') = \frac{\mathbf{D}^a(v) \cdot \mathbf{D}^u(v')}{\|\mathbf{D}^a(v)\| \|\mathbf{D}^u(v')\|}$.

For St Andrews/Facebook, by assuming $\mathcal{M}_s = \{(i, i) | i = 1, 2, \cdots, 6\}$ (which implies $M_0^a = M_0^u = \{1, 2, 3, 4, 5, 6\}$), we can obtain the relative distance similarity scores between the nodes in $V^a \setminus M_0^a$ and the nodes in $V^u \setminus M_0^u$ as shown in Fig.1 (e). From Fig.1 (e), we can observe the following facts. (*i*) Some anonymized nodes (which may be indistinctive with respect to structural similarity), e.g., nodes 14, 19, 23, highly agree with their counterparts and meanwhile disagree with other nodes in the auxiliary graph, which suggests that they can be de-anonymized successfully with a high probability by employing the relative distance similarity based metric. (*ii*) For some nodes, e.g., nodes 11, 21, 26, 27, they are indistinctive on the relative distance similarity with respect to the initial seed selection $\{1, 2, 3, 4, 5, 6\}$. To distinguish them, extra effort is expected, e.g., by utilizing structural similarity collaboratively, employing another seed selection, etc. (*ii*) 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.

Inheritance Similarity. Besides the initial seed mapping, the de-anonymized nodes during each iteration, i.e., $\mathcal{M}_k$, could provide further knowledge when de-anonymize $\Lambda^\delta(M_k^a)$. Therefore, for $v \in \Lambda^\delta(M_k^a)$ and $v' \in \Lambda^\delta(M_k^u)$, we define the knowledge provided by the currently mapped results as the *inheritance similarity*, denoted by $s_I(v, v')$. Formally, $s_I(v, v')$ can be quantified as $s_I(v, v') = \frac{C}{|N_k(v,v')|} \cdot (1 - \frac{|\Delta_v^a - \Delta_{v'}^u|}{\max\{\Delta_v^a, \Delta_{v'}^u\}}) \cdot \sum_{(x,x') \in N_k(v,v')} s(x, x')$ if $N_k(v, v') \neq \emptyset$, and $s_I(v, v') = 0$, otherwise, 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 that (*i*) if two nodes have more common neighbors which have been mapped, then their inheritance similarity

score is high; (*ii*) we also count the degree similarity in defining $s_I(v, v')$. If the degree difference between v and v' is small, then a large weight is given to the inheritance similarity; otherwise, a small weight is given; and (*iii*) we involve the similarity loss in counting $s_I(v, v')$, which implies the inheritance similarity is decreasing with the distance increasing (iteration increasing) between (v, v') and the original seed mapping.

Now, for St Andrews/Facebook, if we assume half of the nodes have been mapped (the first half according to the ID increasing order), then the inheritance similarity between the rest of the nodes in the anonymized graph and the auxiliary graph is shown in Fig.1 (f). From the result, we can observe that under the half number of nodes been mapped assumption, some nodes, e.g., nodes 16, 19, 24, agree with their counterparts and meanwhile disagree with all the other nodes significantly in the auxiliary graph, which implies that they are potentially easier to be de-anonymized when inheritance similarity is taken as a metric. Note that, in Fig.1 (f), we just randomly assume that the known mapping nodes are the first half nodes in the anonymized graph and auxiliary graph. Actually, the accuracy performance of the inheritance similarity measurement could be improved. This is because there are no necessary correlations among the randomly chosen mapping nodes in Fig.1 (f). 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 use of the inheritance similarity in practical de-anonymizaiton.

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 St Andrew, which implies they can be potentially de-anonymized solely based on the structural similarity. On the other hand, for some nodes, due to lacking of distinct topological characteristics, the structural similarity based method can only achieve coarse granularity de-anonymization. Nevertheless and fortunately (from the view of adversary), they may become significantly distinguishable with the knowledge of a small amount of auxiliary information, e.g., nodes 14, 19, and 23 in St Andrews are potentially easy to de-anonymize based on relative distance similarity. In summary, the analysis on real data sets suggests to us to define a unified measurement to properly involve multiple similarity metrics for effective de-anonymization. To this end, we define a *Unified Similarity* (US) measurement by considering the structural similarity, relative distance similarity, and inheritance similarity synthetically for $v \in \Lambda^{\delta}(M_k^a)$ and $v' \in \Lambda^{\delta}(M_k^u)$ in the k-th iteration of our de-anonymization framework as $s(v, v') = c_S \cdot s_S(v, v') + c_D \cdot s_D(v, v') + c_I \cdot s_I(v, v')$, where $c_S, c_D, c_I \in [0, 1]$ are constant values indicating the weights of structural similarity, relative distance similarity, and inheritance similarity, respectively, and $c_S + c_D + c_I = 1$. In addition, we define $s(v, v') = 1$ if $(v, v') \in \mathcal{M}_s$. Now, we are ready to present our US based **D**e-**A**nonymization (DA) framework, which is shown in Algorithm 1.

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

$\mathcal{M}_0 = \mathcal{M}_s$, $k = 0$, *flag* = **true**;
while *flag* = **true do**
  calculate $\Lambda^\delta(M_k^a)$ and $\Lambda^\delta(M_k^u)$;
  if $\Lambda^\delta(M_k^a) = \emptyset$ or $\Lambda^\delta(M_k^u) = \emptyset$, output $\mathcal{M}_k$, **break**;
  for $\forall v \in \Lambda^\delta(M_k^a)$ and $\forall v' \in \Lambda^\delta(M_k^u)$, calculate $s(v, v')$;
  construct a weighted bipartite graph $B_k = (\Lambda^\delta(M_k^a) \cup \Lambda^\delta(M_k^u), E_k^b, W_k^b)$;
  find a *maximum weighted bipartite matching* $\mathcal{M}'$ of B_k;
  for every $(x, x') \in \mathcal{M}'$, if $s(x, x') < \theta$, $\mathcal{M}' = \mathcal{M}' \setminus \{(x, x')\}$;
  let $K = \max\{1, \lceil |\epsilon \cdot \mathcal{M}'| \rceil\}$ and for $\forall (x, x') \in \mathcal{M}'$, **if** $s(x, x')$ *is not the* Top-K *mapping score in* $\mathcal{M}'$ **then**
    $\mathcal{M}' = \mathcal{M}' \setminus \{(x, x')\}$;
  if $\mathcal{M}' = \emptyset$, output $\mathcal{M}_k$ and **break**;
  $\mathcal{M}_{k+1} = \mathcal{M}_k \cup \mathcal{M}'$, 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 k-th iteration, where $E_k^b = \{l_{v,v'}^b | \forall v \in \Lambda^\delta(M_k^a), \forall v' \in \Lambda^\delta(M_k^u)\}$, and $W_k^b = \{w_{v,v'}^b\}$ is the set of all the possible weights on the links in E_k^b. Here, for $\forall (v, v') \in E_k^b$, the weight on this link is defined as the US score between the associated two nodes, i.e., $w_{v,v'}^b = s(v, v')$. 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 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 and starts the iteration procedure. During each iteration, the intended de-anonymizing nodes are calculated first based on the mappings obtained in the previous iteration followed by calculating the US scores between nodes in $\Lambda^\delta(M_k^a)$ and nodes in $\Lambda^\delta(M_k^u)$. Subsequently, based on the obtained US scores, a weighted bipartite graph is constructed between nodes in $\Lambda^\delta(M_k^a)$ and nodes in $\Lambda^\delta(M_k^u)$. Then, we compute a *maximum weighted bipartite matching* $\mathcal{M}'$ on the constructed bipartite graph. To improve the de-anonymization accuracy, we apply two important rules to refine $\mathcal{M}'$: (i) by defining a *de-anonymization threshold* θ, we eliminate the mappings with low US scores in $\mathcal{M}'$. This is because we are not confident to take the mappings with low US scores ($< \theta$) as correct de-anonymizaiton, and more improtantly, they may be more accurately de-anonymized in the following iterations by utilizing confident mapping information obtained in this iteration (this can be achieved since we involve inheritance similarity in the US definition); and (ii) we introduce a *mapping control factor* ϵ, or K equivalently, to limit the maximum number of mappings been accepted as correct de-anonymization. During each iteration, only K mappings with highest US scores will be taken as correct de-anonymization with confidence even if more mappings having US scores greater than the de-anonymizaiton threshold. This strategy has two

benefits. On one hand, only highly confident mappings are kept, which could improve the de-anonymization accuracy. On the other hand, for the mappings been rejected, again, they may be better re-de-anonymized in the following iterations by utilizing the more confident knowledge of the Top-K mappings from this iteration.

Time and Space Complexities Analysis. Let $n = \max\{|V^a|, |V^u|\}$ and $m = \max\{|E^a|, |E^u|\}$. Then, according to combinatorial analysis, Algorithm 1's time complexity is $O(n^2 \log n + mn)$ and space complexity is $O(\min\{n^2, m+n\})$.

4 Generalized Scalable De-anonymization

De-anonymization on Data Sets without Knowledge of Overlap Size. One predicament in practical de-anonymization, which is omitted in existing de-anonymization attacks, is that we do not actually know how large the overlap between the anonymized data and the auxiliary data even we have a lot of auxiliary information available. Therefore, it is unadvisable to do de-anonymization based on the entire anonymized and auxiliary graphs directly, which might cause low de-anonymization accuracy as well as high computational overhead.

To address the aforementioned predicament, guarantee the accuracy of DA, and simultaneously improve de-anonymization efficiency and scalability, we extend DA to an ***A**daptive **De**-**A**nonymization* framework, denoted by ADA. ADA adaptively de-anonymizes G^a starting from a *Core Matching Subgraph* (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, i.e., $G_c^a = G^a[V_c^a]$. 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 $G_c^u = G^a[V_c^u]$.

The CMS is generally defined for two purposes. First, we can employ a CMS to adaptively and roughly estimate the overlap between G^a and G^u in terms of the seed mapping information. On the other hand, we propose to start the de-anonymization from the CMSs, 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.

Now, based on CMS, we discuss ADA as shown in Algorithm 2. In Algorithm 2, μ is the *adaptive factor* which controls the spanning size of the CMS during each adaptive iteration. The basic idea of ADA is as follows. We start the de-anonymization from CMSs G_c^a and G_c^u by running DA. 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 n-

Algorithm 2: **A**daptive **D**e-**A**nonymization (ADA)

generate G_c^a and G_c^u and run DA for G_c^a and G_c^u;
if *Step 1 is ended on the condition that* $\Lambda^\delta(M_k^a) = \emptyset$ *or* $\Lambda^\delta(M_k^u) = \emptyset$ **then**
  if $\Lambda^\mu(V_c^a) = \emptyset$ or $\Lambda^\mu(V_c^u) = \emptyset$, **return**;
  $V_c^a = V_c^a \cup \Lambda^\mu(V_c^a)$, $V_c^a = V_c^a \cup \Lambda^\mu(V_c^a)$;
  $G_c^a = G^a[V_c^a]$, $G_c^u = G^u[V_c^u]$;
  go to Step 1 to de-anonymize unmapped nodes in updated G_c^a and G_c^u;

odes in $\Lambda^\mu(V_c^a)/\Lambda^\mu(V_c^u)$ and repeat the de-anonymization for unmapped nodes. Same as DA, the time and space complexities of ADA are $O(n^2 \log n + mn)$ and $O(\min\{n^2, m+n\})$, respectively.

Disconnected Data Sets. 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 $\bigcup_{i=1}^{m} G_i^a$ and $\bigcup_{j=1}^{n} G_j^u$ 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.

5 Experiments

In this section, we examine the performance of the presented de-anonymization attack on real data sets[3]. In each group of experiments, we specify the employed setup and provide comprehensive analysis. The default settings are: $\alpha = 1.5$, $C = 0.9$, $c_S = 0.2$, $c_D = 0.6$, $c_I = 0.2$, $\theta = 0.6$, $\delta \in \{1, 2\}$, $\mu \in \{1, 2, 3\}$, $\epsilon = 0.5$ and seed number = 5.

Data Sets. In this paper, we employ six well known data sets to examine the effectiveness of the designed de-anonymization framework[4]: St Andrews/Facebook [11][3], Infocom06/DBLP [13][3], Smallbule/Facebook [12][3], ArnetMiner [14], Google+ [10], and Facebook [15]. St Andrews, Infocom06, and Smallbule are three mobility trace data sets. An overview of the three mobility traces is shown in Table 1. We employ the same techniques as in [3] to preprocess the three mobility trace data sets to obtain three anonymized data graphs. To de-anonymize the three anonymized mobility data traces, we employ three auxiliary social network data sets [3] associated with these three mobility traces. For St Andrews, we have a Facebook data set indicating the "friend" relationships among the T-mote users in the trace. For Infocom06, we employ a coauthor data set consisting of 616 authors obtained from DBLP which indicates the "coauthor" relationships among all the attendees of INFOCOM 2005. For Smallblue, we

[3] Due to the space limitation, we put the detailed experimental settings and more results in the Technical Report [17] of this paper.

[4] Not that it has been shown that the classical mobility traces of the (latitude, longitude, timestamp) form can also be represented by graph models [16].

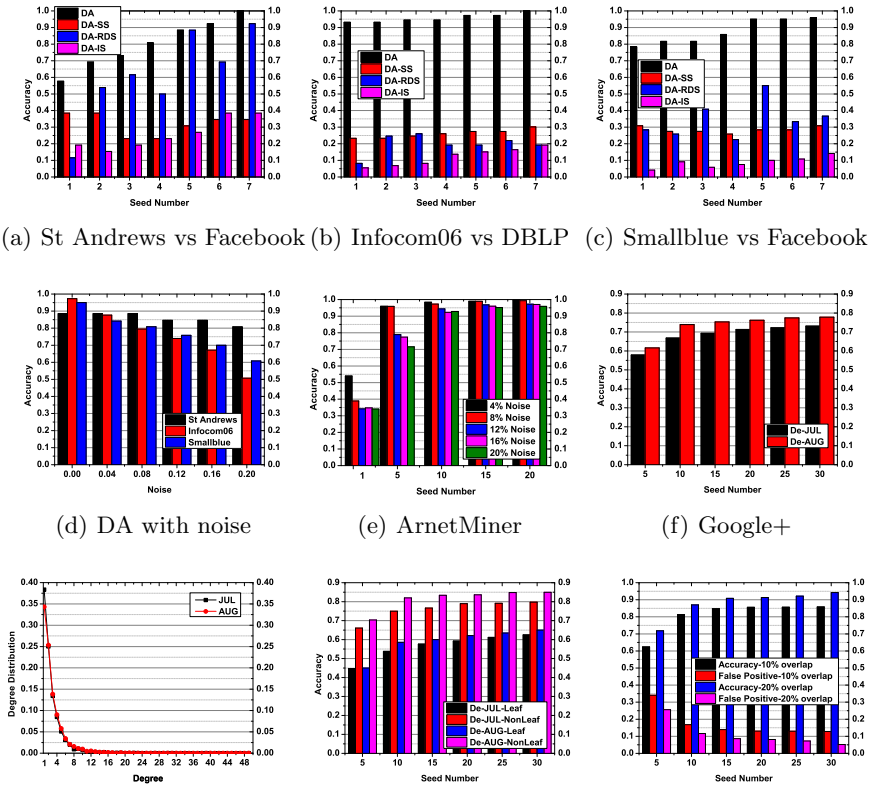
Table 1. Mobility traces.

	St Andrews	**Infocom06**	**Smallblue**
Comm. network type	WiFi	Bluetooth	IM
Comm. nodes No.	27	78	125
Contacts No.	18,241	182,951	240,665
Social network type	Facebook	DBLP	Facebook
Social nodes No.	27	616	400

have a Facebook network among 400 employees from the same enterprise as Smallblue. Note that, the social network data sets corresponding to Infocom06 and Smallblue are supersets of them with respect to involved users.

We also apply the presented de-anonymization attack to social data sets: ArnetMiner [14], Google+ [10], and Facebook [15]. ArnetMiner is an online academic social network, 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. As a new social network, Google+ was launched in early July 2011. We use two Google+ data sets which were created on July 19 and August 6 in 2011 [10], 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 [10], some connections appeared in AUG may not appear in JUL and vise versa. This is because a user may add new connections or disable existing connections. Furthermore, the two data sets 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 data set consists of 63,731 users and 1,269,502 "friend" relationships (links). To use this data set to examine the presented de-anonymization attack, we will preprocess it based on the known hand labeled ground truth.

De-anonymize Mobility Traces. By utilizing the corresponding social networks as auxiliary information, we exploit the presented de-anonymization algorithm DA to de-anonymize the three well known mobility traces St Andrews, Infocom06, and Smallblue. The results are shown in Fig.2 (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, relative distance similarity solely, and inheritance similarity solely, respectively. From Fig.2 (a)-(c), we can see that (*i*) the presented de-aonymization framework is very effective even with a small amount of auxiliary information. For instance, DA can successfully de-anonymize 93.2% of the Infocom06 data just with the knowledge of one seed mapping. For St Andrews and Smallblue, DA can also achieve accuracy of 57.7% and 78.3% respectively with one seed mapping. Furthermore, DA can successfully de-anonymize all the data in St Andrews and Smallblue and 96% of the data of Smallblue with the knowledge of 7 seed mappings; and (*ii*) the US-based de-anonymization is much more effective and stable than structural, relative distance, or inheritance similarity solely based de-anonymization. The reason is that US tries to distinguish a node from



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

(d) DA with noise (e) ArnetMiner (f) Google+

(g) Degree dist. of Google+ (h) Classification (Google+) (i) Facebook

Fig. 2. De-anonymize mobility traces.

multiple perspectives, which is more efficient and comprehensive. As the analysis shown in Section 3, the nodes can be easily differentiated with respect to one measurement but might be indistinguishable with respect to another measurement. Consequently, synthetically characterizing a node as in US is more powerful and stable.

We also examine the robustness of the presented de-anonymization attack to noise and the result is shown in Fig.2 (d) (on the knowledge of 5 seed mappings). In the experiment, we only add noise to the anonymized data. According to the same argument in [2], the noise in the auxiliary data can be counted as noise in the anonymized data. To add p percent of noise to the anonymized data, we randomly add $\frac{p}{2} \cdot |E^a|$ spurious connections to and meanwhile delete $\frac{p}{2} \cdot |E^a|$ existing connections from the anonymized graph (a node may become *isolated* after the noise adding process). For instance, in Fig.2 (d), 20% of noise implies we add 10% spurious connections and delete 10% existing connections of $|E^a|$ from the anonymized data. From Fig.2 (d), we can see that the presented de-anonymization framework is robust to noise. Even if we change 20% of the connections in the anonymized data, the achieved accuracies on St Andrews, Infocom06, and Smallblue are still 80.8%, 50.7%, and 60.8%, respectively. Note

that, when 20% of the connections have been changed, the structure of the anonymized data is significantly changed. In practical, if the anonymized data release is initially for research purposes, this structural change may make the data useless. However, by considering multiple perspectives to distinguish a node, the anonymized data can still be de-anonymized as shown in Fig.2 (d), which confirms the assertion in [2] that data set structure change may not provide effective privacy protection.

De-anonymize ArnetMiner. ArnetMiner can be modeled by a weighted graph where the weight on each relationship indicates the number of coauthored papers by the two authors. To examine the de-anonymization framework, we first anaonymize ArnetMiner by adding p percent noise as explained in the previous 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 Fig.2 (e).

From Fig.2 (e), we can observe that the presented de-anonymization framework is very effective on weighted data. With only knowledge of one seed mapping, more than a half (53.9%) and one-third (34.1%) of the authors can be de-anonymized even with noise levels of 4% and 20%, respectively. Furthermore, when adding 20% of noise to the anonymized data, the presented de-anonymization framework achieves 71.5% accuracy if 5 seed mappings are available and 92.8% accuracy if 10 seed mappings are available; (*ii*) the presented de-anonymization framework is robust to noise on weighted data. When we have 10 or more seed mappings, the accuracy degradation of our de-anonymization algorithm is small even with more noise, e.g., the accuracy is degraded from 99.7% in the 4%-noise case to 96% in the 20%-noise case; and (*iii*) if the available number of seed mappings is 10, the knowledge brought by more seed mappings cannot improve the de-anonymization accuracy significantly. This is because the achieved accuracy on the knowledge of 10 seed mappings is already about 95%. Therefore, to de-anonymize a data set, it is not necessary to spend efforts to obtain a lot of seed mappings. As in this case, to de-anonymize most of the authors, 5 to 10 seed mappings is sufficient.

De-anonymize Google+. Now, we validate the presented de-anonymization framework on the two Google+ data sets JUL and AUG. We first utilize AUG as auxiliary data to deanonymize 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 is shown in Fig.2 (f). Again, from Fig.2 (f), we can see that the presented de-anonymization framework is very effective. Just based on the knowledge of 5 seed mappings, 57.9% of the users in JUL and 61.6% of the users in AUG can be successfully deanonymized. When 10 seed mappings are available, the de-anonymization accuracy can be improved to 66.8% on JUL and 73.9% on AUG, respectively.

However, we also have two other interesting observations from Fig.2 (f): (*i*) when the number of available seed mappings is above 10, the performance im-

provement is not as significant as on previous data sets (e.g., mobility traces, ArnetMiner) even the de-anonymization accuracy is around 70% for JUL and 75% for AUG; and (*ii*) De-AUG has a better accuracy than De-JUL, which implies that the AUG data set is easier to de-anonymize than the JUL data set. To explain the two observations, we assert this is because of the structural property of the two data sets. Follow this direction, we investigate the degree distribution of JUL and AUG as shown in Fig.2 (g). From Fig.2 (g), we can see that the degree of both JUL and AUG generally follows a *heavy-tailed distribution*. In particular, 38.4% of the users in JUL and 34.3% of the users in AUG have degree of one, named *leaf users*. This is normal since Google+ was launched in early July 2011, and JUL and AUG are data sets crawled in July and August of 2011, respectively. That is also why JUL has more leaf users than AUG (a user connects more people later). Now, we argue that the leaf users cause the difficulty in improving the de-anonymization accuracy. From the perspective of graph theory, the leaf users limit not only the performance of our de-anonymization framework but also the performance of any de-anonymization algorithm. An explanatory example is as follows. Suppose $v \in V^a$ is successfully de-anonymized to $v' \in V^u$. In addition, the two neighbors x and y of v and the two neighbors x' and y' of v' are all leaf users. Then, even $x' = \gamma(x)$, $y' = \gamma(y)$, and v has been successfully de-anonymized to v', it is still difficult to make a decision to map x (or y) to x' or y' since $s(x, x') \approx s(x, y')$ from the view of graph theory. Consequently, to accurately distinguish x, further knowledge such as semantic information is required.

To support our argument, we take an insightful look on 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 one or *non-leaf user set* if its degree is greater than one. Then, we re-calculate the de-anonymization accuracy for leaf users and non-leaf users and the results are shown in Fig.2 (h), where De-JUL-Leaf/De-AUG-Leaf represents the ratio of leaf nodes that have been successfully de-anonymized in JUL/AUG while De-JUL-NonLeaf/De-AUG-NonLeaf represents the ratio of non-leaf users that have been successfully de-anonymized in JUL/AUG. From Fig.2 (h), we can see that (*i*) the successful de-anonymization ratio on non-leaf users is higher than that on leaf users in JUL and AUG. This is because non-leaf users carry more structural information; and (*ii*) considering the results shown in Fig.2 (f), the de-anonymization accuracy on non-leaf users is higher than the overall accuracy and the de-anonymization accuracy on leaf users is lower than the overall accuracy. The two observations on Fig.2 (h) confirms our argument that leaf users are more difficult than non-leaf users to de-anonymize. Furthermore, this is also why De-AUG has higher accuracy than De-JUL in Fig.2 (f). AUG is easier to de-anonymize since it has less leaf users than JUL.

De-anonymize Facebook. Finally, we examine ADA on Facebook. Based on the hand labeled ground truth, we partition the data sets into two about-equal parts utilizing the method employed in [2], 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, the achievable accuracy and the induced false positive error of ADA are shown in Fig.2 (i). As a fact, most of the existing de-anonymization attacks are not very effective for the scenario that the overlap between the anaonymized data and the auxiliary data is small or even cannot work totally. Surprisingly, for ADA, we can observe from Fig.2 (i) that (*i*) based on the proposed CMS, ADA can successfully de-anonymize 62.4% of the common users with false positive error of 34.1% when the overlap is 10% and 71.8% of the common users with false positive error of 25.6% when the overlap is 20% with the knowledge of just 5 seed mappings; (*ii*) the de-anonymization accuracy is improved to 81.3% (resp., 85.6%) and the false positive error is decreased to 16.8% (resp., 13%) when the overlap is 10% and 10 (resp., 20) seed mappings available, and the de-anonymization accuracy is improved to 87% (resp., 90.8%) and the false positive error is decreased to 11.6% (resp., 8.6%) when the overlap is 20% and 10 (resp., 20) seed mappings available, which demonstrates that ADA is very effective in dealing with the partial data overlap situation; and (*iii*) ADA has a higher de-anonymization accuracy and lower false positive error in the 20% data overlap scenario than that in the 10% data overlap scenario. This is because a larger overlap size implies a common node will carry much more similar structural information in both graphs, and thus it can be de-anonymized with higher probability and accuracy. From Fig.2 (i), we can also see that 10 seed mappings are sufficient to achieve high de-anonymization accuracy and low false positive error. Therefore, ADA is applicable with efficiency and performance guarantee in practical.

6 Conclusion

In this paper, we present a novel and effective de-anonymization attack based on a Unified Similarity (US) measurement which synthetically incorporates multiple data structural factors. The experimental results demonstrate that the presented de-anonymization framework is very effective and robust to noise.

Acknowledgments

Mudhakar Srivatsa's research was sponsored by 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 US 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 the Interdisciplinary Research Opportunities (IDROP) Program.

References

1. L. Backstrom, C. Dwork, and J. Kleinberg, *Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, WWW 2007.
2. A. Narayanan and V. Shmatikov, *De-anonymizing Social Networks*, S&P 2009.
3. M. Srivatsa and M. Hicks, *Deanonymizing Mobility Traces: Using Social Networks as a Side-Channel*, CCS 2012.
4. A. Narayanan and V. Shmatikov, *Robust De-anonymization of Large Sparse Datasets (De-anonymizing the Netflix Prize Dataset)*, S&P 2008.
5. D. Goodin, *Poorly anonymized logs reveal NYC cab drivers detailed whereabouts*, *http://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/*.
6. K. Singh, S. Bhola, and W. Lee, *xBook: Redesigning Privacy Control in Social Networking Pl atforms*, USENIX 2009.
7. P. Hornyack, S. Han, J. Jung, S. Schechter, and D. Wetherall, *"These Aren't the Droids You're Looking For": Retrofitting Android to Protect Data from Imperious Applications*, CCS 2011.
8. M. Egele, C. Kruegel, E. Kirda, and G. Vigna, *PiOS: Detecting Privacy Leaks in iOS Applications*, NDSS 2011.
9. T. Opsahl, F. Agneessens, and J. Skvoretz, *Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths*, Social Networks, Vol 32, pp. 245-251, 2010.
10. N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. C. R. Shin, E. Stefanov, E. Shi, and D. Song, *Jointly Predicting Links and Inferring Attributes using a Social-Attribute Network (SAN)*, SNA-KDD 2012.
11. G. Bigwood, D. Rehunathan, M. Bateman, T. Henderson, and S. Bhatti, *CRAWDAD data set st_andrews/sassy (v. 2011-06-03)*, Downloaded from *http://crawdad.cs.dartmouth.edu/∼crawdad/st_andrews/sassy/*, June 2011.
12. Smallblue, *http://domino.research.ibm.com/comm/research_projects.nsf/pages/smallblue.index.html* .
13. J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau, *CRAWDAD data set cambridge/haggle (v. 2009-05-29)*, Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle, May 2009.
14. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, *ArnetMiner: Extraction and Mining of Academic Social Networks*, KDD 2008.
15. B. Viswanath, A.n Mislove, M. Cha, and K. P. Gummadi, *On the Evolution of User Interaction in Facebook*, WOSN 2009.
16. H. Pham, C. Shahabi, and Y. Liu, *EBM - An Entropy-Based Model to Infer Social Strength from Spatiotemporal Data*, Sigmod 2013.
17. S. Ji, W. Li, M. Srivatsa, J. He, and R. Beyah, *Technical Report: Data De-anonymization: From Mobility Traces to On-line Social Networks*, http://users.ece.gatech.edu/∼sji/Paper/isc14TechReport.pdf.