Finding the Most Appropriate Auxiliary Data for Social Graph Deanonymization

Priya Govindan Sucheta Soundarajan Tina Eliassi-Rad
Rutgers University
{priyagn, s.soundarajan, eliassi}@cs.rutgers.edu

ABSTRACT

Given only a handful of local structural features about the nodes of an anonymized social graph, how can an adversary select an auxiliary (a.k.a. non-anonymized, known) graph to help him/her deanonymize (a.k.a. re-identify) the individuals in the graph? Examples of local structural features are node's degree, node's clustering coefficient, edge density of the node's neighbors, *etc.* The objective of the adversary is to find an auxiliary graph that has the maximum node-overlap with the anonymized graph. We present conditions under which an adversary may *estimate* the node-overlap between the graphs; and thus be able to pick the most appropriate auxiliary graph. Specifically, we consider two scenarios. In the first scenario, the adversary has no information about the anonymized graph. We call this situation the *no seeds* case. In the second scenario, the adversary is able to gain some information about the anonymized graph. For example, the adversary is able to find out that a handful of individuals are *present* in the anonymized graph. We call this scenario the *some seeds* case. Our findings indicate that (1) in the *no seeds* case, an adversary can predict when the node-overlap between the anonymized and auxiliary graphs is low; (2) in the *some seeds* case, an adversary can identify pairs of anonymized and auxiliary graphs with high node-overlap; and (3) in the *some seeds* case, an adversary can effectively *learn* to predict the node-overlap on different auxiliary graphs.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; E.1 [**Data Structures**]: Graphs and Networks; K.4.1 [**Public Policy Issues**]: Privacy

General Terms

Algorithms, Design, Performance, Experimentation

KDD'14 Data Ethics Workshop New York, NY USA

Keywords

Social graphs, deanonymization, re-identification

1. INTRODUCTION

Suppose an organization (such as the National Institute of Health) wishes to release social graph data.[1] To protect the privacy of the individuals in the data, the organization first anonymizes (a.k.a. de-identifies) the graph data. As a further layer of privacy, rather than releasing the topology of the graph itself (i.e., its adjacency matrix), the organization releases data in the form of a feature matrix describing local structural properties of each node (e.g., degree, clustering coefficient, average degree of neighbors, *etc*).

An adversary interested in re-identifying individuals in the anonymized graph will seek an auxiliary graph in which node-identities are known. The adversary has many possible auxiliary graphs from which to choose. The question is: *which auxiliary graph is the most appropriate choice*? To break the privacy of the greatest number of individuals, the adversary should select the auxiliary graph that has the greatest node-overlap with the anonymized network. Thus, the problem of picking the most appropriate auxiliary network reduces to the following problem: **predict the percentage of nodes from a known auxiliary graph that is present in an anonymized graph**—i.e., predict the *node-overlap* between a pair of graphs where one is known and the other is anonymized. Recall that the only data available from the anonymized graph is a feature matrix whose rows are the anonymized nodes and whose columns are local structural features (see Section 2 for details). This problem is challenging because:

- Challenge #1: We are only given a feature matrix representing local structural features of nodes in the anonymized network. We do not know which nodes are adjacent to other nodes; and thus cannot propagate beliefs about whether nodes are present or absent in the anonymized graph. This also means that we cannot take advantage of the sparsity that is inherent in the topology of social graphs since feature matrices are (by definition) dense.

- Challenge #2: The anonymized graph may be drastically different from the auxiliary graph. In particular, without any side information, it is difficult to deter-

[1]A dating network is an example of a social graph. We use the terms network and graph interchangeably.

mine how much a specific node's structure may vary from one graph to the other.

- Challenge #3: Nodes that are present in both the anonymized and the auxiliary graphs may have different feature values (a.k.a. *profiles*) than nodes that are present in only one graph. But, these profiles are inconsistent across various domains. For example, rules that identify present nodes in the academic coauthorship domain are seldom useful in the Hollywood collaboration domain.

Our empirical study (Section 3) demonstrates that when an adversary is able to overcome these challenges, then he/she may pose a serious threat to individuals' privacy because he/she can select the most appropriate auxiliary data for deanonymization −i.e., a known dataset that has high node-overlap with the anonymized graph.

Our extensive empirical study has three parts. In **part one** (Section 3.2), we illustrate the aforementioned challenges on real networks. Specifically, in some cases, the structural properties of nodes present in both the anonymized and the auxiliary graphs may be very similar to the properties of the nodes that are absent from the auxiliary graph. In other cases, where the present and absent nodes have different properties, these properties are unlikely to be consistent across graphs. In **part two** (Section 3.3), we present a series of results that illustrate when an adversary may be able to overcome the aforementioned challenges. We consider two versions of this problem. In the first version, we assume that nothing is initially known about which nodes in the anonymized feature matrix are present or absent in the auxiliary graph. We call this the *no seeds* case. In this case, we show that an adversary is able to accurately predict when the node-overlap between the auxiliary graph and the anonymized graph is below 20%. The adversary can thus reject an auxiliary graph as a potential candidate for further privacy-breaking endeavors. In the second version, we consider the case when the adversary has somehow managed to learn the identities of some nodes in the anonymized graph. We call this the *some seeds* case. In such a situation, we show that the adversary can break privacy by using these *seed* nodes to estimate each node's structural change from the auxiliary graph to the anonymized graph. We present a method which allows an adversary to use this information to predict the amount of node-overlap between the anonymized and auxiliary data; and show that this predicted node-overlap can be surprisingly accurate given a handful of seeds. In **part three** (Section 3.3), we consider the *some seeds* case again and show that the adversary can effectively use transfer learning to predict the node-overlap between a second auxiliary graph and the same anonymized feature matrix, further defeating privacy.

Our contributions are as follows:

- We reformulate the problem of finding the most appropriate auxiliary data for deanonymizing a social graph's structural feature matrix into the problem of predicting the node-overlap between the auxiliary data and the anonymized structural feature matrix.
- We demonstrate the challenges involved in solving this node-overlap problem, including issues of similarity between the nodes that are present in both graphs and the nodes that are present in only one graph.
- We show that despite the challenges, an adversary can accurately predict when the node-overlap between the anonymized and auxiliary data is low. In addition, given some side information, he/she can accurately predict when the node-overlap is high. Moreover, he/she can use information about the node-overlap between the anonymized graph and an auxiliary graph to predict the node-overlap between the same anonymized graph and a different auxiliary graph.

The outline of our paper is as follows. In Section 2, we elaborate on the problem definition and include a description of our assumptions. Section 3 presents our empirical study. In Section 4, we discuss related works. Section 5 concludes the paper.

2. PROBLEM DEFINITION

We consider several versions of the same basic problem. In all versions, an organization has released a feature matrix F_{anon} containing structural properties of nodes in an anonymized graph G_{anon}.

We assume that the organization releasing F_{anon} first applies an anonymization technique to G_{anon}, then computes the features values that fill F_{anon}. Many anonymization techniques exist (see Section 4). One often selects a technique based on how much utility one wants to preserve. In this paper, we apply an information-theoretic anonymization technique by randomly perturbing a fraction of edges [4]. Specifically, for a given graph $G = (V, E)$ and a fraction $r \in [0, 1]$, we (1) remove unique identifiers on nodes; (2) randomly delete k edges from G, where $k = r \times |E|$; and (3) add k edges to G by randomly picking pairs of unconnected nodes and adding edges between them. This anonymization technique preserves the number of edges in a graph. However, it may not preserve other properties such as degree distribution or clustering coefficient, which makes it harder for an adversary to identify a node based on properties such as degree.

The rows of F_{anon} correspond to nodes in G_{anon}. The columns of F_{anon} are the following seven structural features:

1. Node's degree
2. Average degree of node's neighbors
3. Node's clustering coefficient
4. Average clustering coefficient of node's neighbors
5. Number of edges between node's neighbors
6. Number of nodes adjacent to node's neighbors
7. Number of edges outgoing from the node's neighbors

Berlingerio et al. [2] showed that these seven local structural features correspond to four social theories−namely, Social Capital, Social Exchange, Balance, and Structural Hole. Recall that the organization does not release G_{anon} itself.

The adversary obtains an auxiliary graph G_{aux}. He/she wishes to predict the fraction of nodes in G_{anon} that are also present in G_{aux}. If this fraction is not low, then the adversary knows that G_{aux} is a suitable graph for his/her deanonymization task.

In different versions of this problem, the adversary may have additional *seed* information,[2] which informs the adversary whether a handful of individuals are present or absent between the anonymized and auxiliary graphs. We call these *seed labels*. In particular, a node in the auxiliary graph is labeled *present* if it appears in both the anonymized and auxiliary data; or it is labeled *absent* if it appears in only the auxiliary data. Furthermore, the adversary may be able to match some individuals in the auxiliary graph to individuals in the anonymized data (see Section 4). We call these *seed matches*.

3. EMPIRICAL STUDY

This section is divided into three parts: (i) datasets, (i) challenges, and (iii) methods and results.

3.1 Datasets

Table 1 lists the four datasets used in our experiments. They include two communication graphs and two social graphs. The two communication graphs are *Twitter Retweets* and *Yahoo! IM*. The two social graphs are *DBLP Computer Science Bibliography* and *IMDB Movie Collaborations*.

Twitter Retweets is data from May to September 2009. We extracted graphs from tweets such that an edge between two Twitter users exists if one of the users had retweeted the other user's tweet during that period. We divide the data into five graphs where each graph represents a month of activity.

Yahoo! IM[3] was collected for the duration of 28 days in April 2008. In this dataset, nodes represent users and an edge between two users indicate that they exchanged messages during that period.

DBLP Computer Science Bibliography[4] consists of coauthorship graphs that were collected from 2005 to 2009. We extracted the co-authorship graphs for the papers published in CIKM, ICDM, KDD, SIGMOD, SDM, VLDB.

IMDB Movie Collaborations[5] graphs consists of individuals credited in movies from 1950 to 1955. The individuals are represented by nodes and an edge between two individuals exists if they have both been credited for at least one movie during that period. We break the data into six graphs (1950-1955), such that a graph represents the data collected that year.

3.2 Challenges

The problem of predicting node-overlap between an auxiliary graph G_{aux} and an anonymized feature matrix F_{anon} is challenging for a multitude of reasons. A major challenge is that it is difficult to distinguish between nodes in F_{anon} that are present in G_{aux} and those that are absent from G_{aux}. Thus, one cannot simply calculate the node-overlap between F_{anon} and G_{aux} by predicting individual node labels. To showcase this challenge, we examine the structural characteristics of nodes in G_{aux} that are also present in F_{anon}, and compare these characteristics to the characteristics of nodes in G_{aux} that are absent in F_{anon}. It is reasonable to believe that present and absent nodes may be structurally distinct. If, for example, G_{aux} and F_{anon} consist of two separate daily snapshots of an instant messaging network, it is possible that the higher degree nodes would be more active and thus be more likely to exist (i.e., be present) across multiple snapshots.

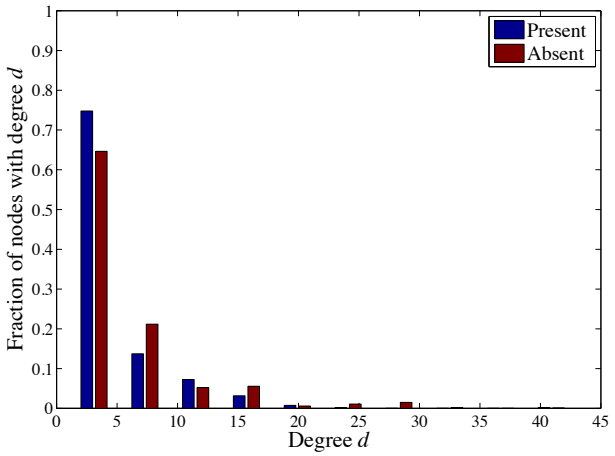

Figure 1: Histogram of node-degree values from the 2005 DBLP auxiliary graph. 'Present' nodes are also present in the anonymized version of the 2006 DBLP graph. 'Absent' nodes are not present in the anonymized version of the 2006 DBLP graph. The observed differences are not sufficient to learn an effective (i.e., better than random) classifier that distinguishes between present and absent nodes.

To examine the aforementioned possibility, we select pairs of graphs from our data (described in Section 3.1). We label one graph as G_{aux} and the other as G_{anon}. We then randomly perturb 5% of the edges in the graph selected as G_{anon};[6] and calculate its structural feature matrix F_{anon}. Subsequently, we annotate nodes into one of two classes: the 'present' class containing nodes from G_{aux} that are also present in F_{anon}, and the 'absent' class containing nodes from G_{aux} that are absent from F_{anon}. Using values from F_{anon}, we calculate distributions of feature values from the present and absent classes. If these classes have significant differences in their structural features, then an adversary may utilize them to break the privacy of a different anonymized graph.

Figure 1 contains the histogram of degree values for each class when the auxiliary graph was a 2005 snapshot of DBLP and the anonymized graph was a perturbed version of a 2006 snapshot of DBLP. We observe that in this case, there are minor differences in the degrees of present and absent nodes. However, these differences are not sufficient for learning a classifier (such as a support vector machine), which can accurately distinguish between present and absent nodes. Indeed, the classifiers performed no better than random.

Even when the differences between present and absent nodes are noticeable, they are inconsistent across various domains. For example, Figure 1 shows that in the DBLP coauthorship domain, a slightly higher percentage of present nodes tend to have very low degrees compared to absent nodes. However, Figure 2 (which compares nodes from a

[2]See [10] for ways in which an adversary may come upon such information.

[3]`http://sandbox.yahoo.com/`

[4]`http://www.informatik.uni-trier.de/~ley/db/`

[5]`http://www.imdb.com/`

[6]See Section 2 for details on anonymization of G_{anon}.

Real-world Graphs	Avg. # of Nodes (Std. Dev.)	Avg. # of Edges (Std. Dev.)	Avg. Node-Overlap Between Graph Pairs (Std. Dev.)
Twitter Retweet Monthly Graphs from May to Sep. 2009	63,232.5 (34,274.6)	81909.9 (51,498.2)	0.42 (0.30)
Yahoo! IM Weekly Graphs in April 2008	84,623.5 (9,435.0)	261,144.2 (48,944.6)	0.91 (0.08)
DBLP Co-authorship Yearly Graphs from 2005 to 2009	2,039.6 (429.3)	4,022.3 (1,095.1)	0.43 (0.29)
IMDB Collaboration Yearly Graphs from 1950 to 1955	10,887.6 (451.8)	236,120.2 (51,611.1)	(0.51) (0.23)

Table 1: Real-world graphs used in our experiments. Each graph was anonymized by perturbing 5% of the edges. The node-overlap values between graph pairs (where a pair refers to the anonymized and the auxiliary data) ranges from a minimum of 0.07 to a maximum of 1 in our datasets. Here we only report the averages and standard deviations.

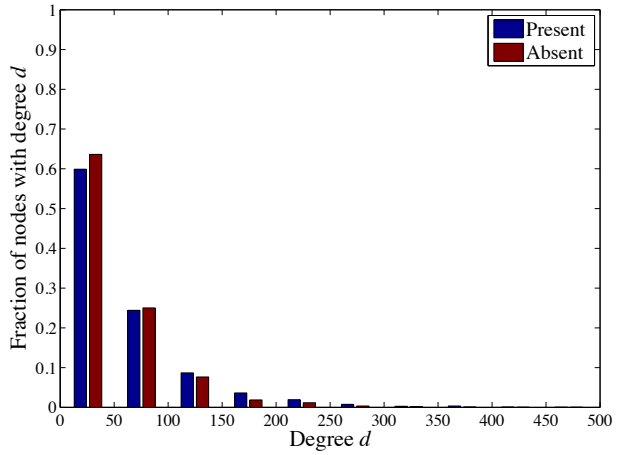

Figure 2: Histogram of node-degree values from the 1950 IMDB auxiliary graph. 'Present' nodes were also present in the anonymized version of the 1951 IMDB graph. 'Absent' nodes were not present in the anonymized version of the 1951 IMDB graph. Note that the differences between the two classes are not the same as the differences shown for the DBLP coauthorship graphs in Figure 1.

1951 snapshot of IMDB to a 1950 snapshot of IMDB) shows the opposite trend, where a slightly higher percentage of absent nodes tend to have very low degrees.

Across the four domains that we considered (academic coauthorships, Hollywood collaborations, instant messaging, and tweets), we did not observe any consistent trends that would allow for an accurate prediction of present vs. absent class labels on individual nodes based solely on local structural features. This lack of consistency makes the problem of predicting node-overlap between the anonymized and auxiliary data challenging.

3.3 Methods and Results

Despite the aforementioned challenges, it is still possible for an adversary to make some headway in selecting the "most appropriate" auxiliary data. Here, we present two results, corresponding to the cases when the adversary has no further information about F_{anon} (i.e., the *no seeds* case; see Section 3.3.1) as well as the case when some supplementary information about nodes in F_{anon} is available (i.e., the *some seeds* case; see Section 3.3.2).

3.3.1 Predicting Node-Overlap in the No Seeds Case

We consider the case where an adversary only has the feature matrices F_{anon} and calculates F_{aux} from G_{aux}. He/she has no other information. As we described in Section 3.2, this is a hard problem. However, we propose a simple method to predict overlap based on the distance between the centers of the feature matrices. We observe that using this method an adversary can confidently identify auxiliary graphs that have low node-overlap with the anonymized data; and thus are not appropriate for deanonymization.

Given F_{aux} and F_{anon}, we estimate their node-overlap as follows:

$$\begin{aligned} \textit{Predicted Overlap} &= \textit{Maximum Overlap} \times \\ &(1 - \textit{Canberra}(\textit{Centroid}(F_{aux}), \textit{Centroid}(F_{anon}))) \end{aligned} \tag{1}$$

where $Centroid(M)$ takes a feature matrix M with k columns and outputs a mean vector of size k. We use the normalized Canberra distance[7] in Equation 1 since it is sensitive around small values. *Maximum Overlap* is the maximum ratio of nodes that can be present in both graphs. It is defined as follows:

$$\textit{Maximum Overlap} = \frac{min(|F_{aux}|, |F_{anon}|)}{|F_{aux}|} \tag{2}$$

$|F_{aux}|$ and $|F_{anon}|$ denote the number of rows (i.e., nodes) in the F_{aux} and F_{anon} matrices, respectively.

We define the *true* node-overlap between F_{aux} and F_{anon} as follows:

$$\textit{True Overlap} = \frac{|F_{aux} \cap F_{anon}|}{|F_{aux}|} \tag{3}$$

$|F_{aux} \cap F_{anon}|$ is the number of nodes that are present in both the anonymized and auxiliary data.

Figure 3 shows that the predicted overlap (defined in Equation 1) increases as the true node-overlap (defined in Equation 3) increases. Note that when the predicted node-overlap

[7] $Canberra(\vec{u}, \vec{v}) = \sum_{i=1}^{d} \frac{|u[i]-v[i]|}{|u[i]|+|v[i]|}$, where d is the dimension of vectors $\vec{u}$ and $\vec{v}$. We divide $Canberra(\vec{u}, \vec{v})$ by d to normalize it.

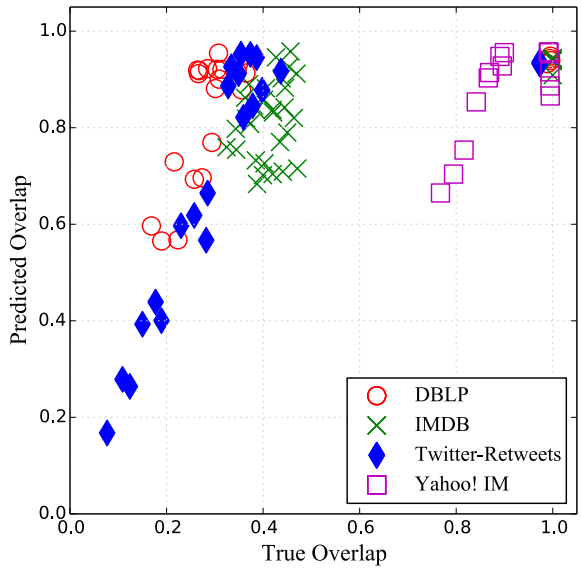

Figure 3: *Predicted overlap* (Eq. 1) vs. *true overlap* (Eq. 3). Graph pairs with a predicted node-overlap of less than 0.5 have low true node-overlap (at most 0.2).

is low (i.e., less than 0.5), then the true node-overlap is at most 0.2. Thus, when an adversary comes across an F_{aux}, for a given F_{anon}, whose predicted node-overlap according to Equation 1 is low, he/she can conclude that such an auxiliary graph will have low node-overlap, discard it, and continue his/her search for a better auxiliary graph.

3.3.2 Predicting Node-Overlap in the Some Seeds Case

In this section, we describe experiments where an adversary first obtains some seed information (either as seed-labels or seed-matches) about the anonymized graph; and then proceeds to estimate the node-overlap between the anonymized and auxiliary data.

Adversary obtains seed-labels. In this case, the adversary has 'present' vs. 'absent' labels on a fraction of nodes in the auxiliary graph and uses this side information to predict the node-overlap between F_{anon} and G_{aux}. We study the effect of varying the fraction of seed-labels on predicting the node-overlap between F_{anon} and G_{aux}; and show that even with as low as 10 seed-labels, an adversary can accurately infer the node-overlap between the anonymized and auxiliary data.

Recall that the adversary has G_{aux}, F_{anon}, and 'present' or 'absent' labels on a fraction of the nodes in G_{aux}. A 'present' label on a node in G_{aux} denotes that the node in G_{aux} is also present in F_{anon}; and an 'absent' label denotes otherwise. An adversary could estimate the node-overlap between G_{aux} and F_{anon} by collecting the labels of the seed nodes uniformly at random for as large a fraction as possible. As shown in Table 1, the node-overlap between the graphs has a wide range. In cases where the node-overlap is low (i.e., the ratio of 'present' to 'absent' labels is skewed), it is difficult to accurately estimate the node-overlap from a small sample of seed nodes.

Figure 4 reports the estimated node-overlap when we have 10 seed-labels (i.e., we know whether 10 nodes in G_{aux} are present or absent in F_{anon}). These nodes were chosen uniformly at random from G_{aux}. Figure 4 shows that an adversary can estimate the overlap with a Mean Absolute Error (MAE) of 0.03 when given only 10 seed-labels. Here, MAE is defined on a set of n anonymized and auxiliary graph pairs as follows:

$$MAE(\{pair_1, \ldots, pair_n\}) = \frac{1}{n}\sum_{i=1}^{n} |True\_Overlap(pair_i) - Predicted\_Overlap(pair_i)| \quad (4)$$

The true overlap is the same as the definition in Equation 3. The predicted overlap here is the number of 'present' seed-labels divided by the total number of seed-labels.

The very low MAE (of 0.03 for only 10 seed-labels) demonstrates the vulnerability of anonymized graphs (released only by their local structural feature matrices) to accurate estimations of node-overlap with auxiliary graphs. We also conducted experiments with 50 and 100 seed-labels, which yielded MAE values of 0.01 and 0.002, respectively. As expected with more seed-labels, the MAE values decrease.

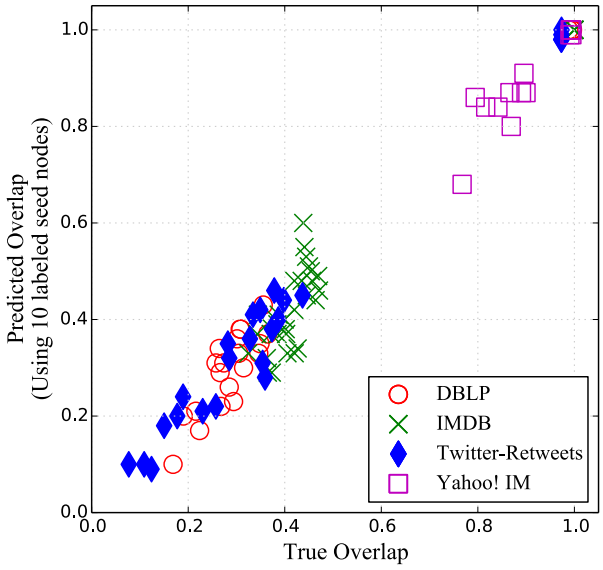

Figure 4: *Predicted overlap* (fraction of 'present' labels in the seed-labels) vs. *true overlap* (Eq. 3). Mean absolute error (MAE defined in text) is 0.03 when using 10 seed-labels. MAE for 50 and 100 seed-labels are 0.01 and 0.002, respectively.

Adversary obtains seed-matches. The aforementioned method requires the adversary to obtain an unbiased sample of seed-labels, which may be difficult in real-world settings. So, we now consider a case where the adversary has obtained a biased sample of *seed-matches* (say for a few vulnerable individuals in G_{aux}). A *seed-match* maps an individual in G_{aux} to a specific row in F_{anon}. This is different than seed-labels, where for a seed individual in G_{aux} we only know whether it is present or absent in F_{anon}. Now that we have seed-matches, we can define the concept of *lookalikes*.

We define *lookalikes* for a node x in G_{aux} as the fraction of nodes in F_{anon} that are at least as similar to x as its

matched node in F_{anon}. Specifically, given G_{aux}, we create a feature matrix F_{aux} by computing the same local structural properties as released with F_{anon} on the nodes in G_{aux}. Then for each node x in F_{aux}, we compute the fraction of nodes in F_{anon} whose structural feature vector (in terms of Canberra distance) is less than or equal to the structural feature vector of x than its matched node in F_{anon}. The value for lookalikes between F_{anon} and G_{aux} is the average of lookalikes between the nodes in G_{aux} to the nodes in F_{anon}. Intuitively, lookalikes can be thought of as a measure of structural change between nodes from one graph to another. A low value for lookalikes means that there is little structural change between F_{anon} and G_{aux}.

We conduct an experiment where we *predict* the value of lookalikes for an ⟨anonymized, auxiliary⟩ graph pair, by estimating it from a set of seed-matches (i.e., 'present' nodes whose matched node in F_{anon} is known). Note that in the previous experiment, we required the nodes with seed-labels to be picked at random. Whereas in this experiment, we do not need to make such an assumption. An adversary might be able to carefully pick vulnerable nodes for whom the seed-matches can be easily found. We assume that the adversary seeks matches for p% of the nodes in F_{aux}.[8] Of the p% of nodes in F_{aux} for which the adversary seeks a match, it is possible that matches may be found for fewer than p% of nodes. We use the average lookalikes of the seed-matches (which were found) as a predicted estimate of the lookalikes for G_{aux}.

Using the seed nodes (which may include 'absent' nodes), we can also predict the node-overlap. Specifically, we train an AdaBoost classifier on the seed nodes from F_{aux} and their labels; then predict labels on the remaining nodes in F_{aux}. Here, node-overlap is estimated as the ratio of nodes in F_{aux} predicted to be 'present' in F_{anon} by the AdaBoost classifier.

Figure 5 shows the predicted *lookalikes* compared to the predicted node-overlap. The predicted *lookalikes* are the averages over 10 runs for each graph pair. In each run, a random sample of p =10% of the nodes in G_{aux} were chosen as seeds. The predicted node-overlap are the result of ten-fold cross-validation to train and test an AdaBoost classifier. The same seeds were used to predict lookalikes and node-overlap. We observe a cluster in the top left portion of the plot; these are the cases where the *lookalikes* value is low, and the predicted node-overlap is very high. A closer examination of these points reveals that the true node-overlap in these cases is also very high. Recall that, *lookalikes* is a measure of structural change between nodes from one graph to another; hence this figure illustrates that when the anonymized and auxiliary graphs are structurally similar, and their predicted node-overlap value is high, then their true node-overlap is also likely to be high. By using this method, an adversary can confirm that his/her auxiliary graph is appropriate to deanonymize a given anonymized graph.

Adversary performs transfer learning. We also investigate the following transfer-learning scenario. We assume the adversary has the anonymized data F_{anon}, an auxiliary graph $G1_{aux}$, and seed-labels ('present' vs. 'absent') on *all* nodes in $G1_{aux}$. The adversary uses one of the aforementioned methods and discovers that the node-overlap between F_{anon} and $G1_{aux}$ is low. Instead of discarding $G1_{aux}$ (for which he/she has seed-labels), the adversary obtains another auxiliary graph $G2_{aux}$ that has low node-overlap with $G1_{aux}$. Now, his/her task is to calculate the node-overlap between F_{anon} and $G2_{aux}$. Next, we describe how an adversary can use the labels from $G1_{aux}$ to better estimate the node-overlap between F_{anon} and $G2_{aux}$.

[8]Predicting lookalikes is a harder problem than predicting node-overlap, which is why an adversary would need p% of nodes as seed-matches, rather than a few seed-labels.

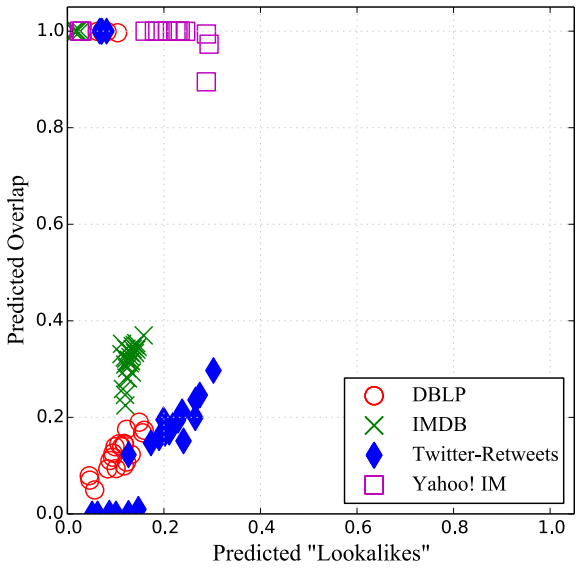

Figure 5: *Predicted overlap* vs. *predicted lookalikes*: Using vulnerable seed-matches (here 10% of the nodes in G_{aux}), an adversary can predict very high (close to 100%) predicted node-overlap when the predicted lookalikes is less than 0.1. See text for the definitions of predicted overlap and predicted lookalikes.

In our transfer-learning experiments, we first create structural feature matrices $F1_{aux}$ and $F2_{aux}$ from $G1_{aux}$ and $G2_{aux}$, respectively. These feature matrices have the same structural features as F_{anon} (see Section 2). We then randomly select 10% of the nodes in $G1_{aux}$ (i.e., rows in $F1_{aux}$) as our training set and train an AdaBoost classifier. Recall that we have seed-labels on all of the nodes in $G1_{aux}$. Using the trained classifier, we predict labels on the nodes in $G2_{aux}$ (i.e., rows in $F2_{aux}$). Here, node-overlap is estimated as the ratio of nodes in $G2_{aux}$ predicted to be 'present' in F_{anon}. Figure 6 plots the average values across 10 runs of this experiment; and shows that the predicted node-overlap is better here (in the tranfer-learning case) than the case with no seeds (as shown in Figure 3). These experiments illustrates that if an adversary has prior knowledge in terms of an auxiliary graph with labels, he/she is better equipped to estimate node-overlap with other graphs. In such cases, an adversary can choose other auxiliary graphs with higher node-overlaps that can aid him/her in deanonymizing F_{anon}.

4. RELATED WORK

Related work can be divided into (1) anonymization techniques for networks and (2) techniques for deanonymizing a social graph given its anonymized adjacency matrix.

Anonymization techniques exist that preserve the privacy

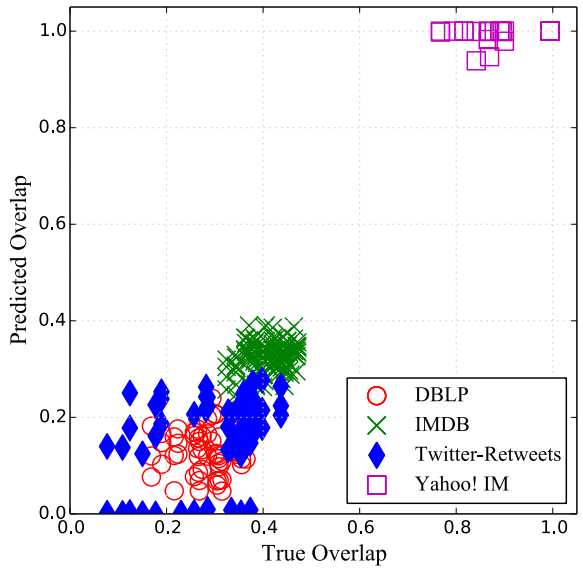

Figure 6: *Predicted overlap* (by transfer-learning) vs. *true overlap* (Eq. 3 on $G2_{aux}$). Most values lie on the diagonal (with root mean square error of 0.14 from the 45° line) indicating that transfer learning predictions are good estimates of the true overlap.

of the data while providing the intended utility. Hay et al. [6] show that using structural queries an adversary could identify nodes. They propose a method to anonymize the graph by grouping the nodes before releasing it. In another work, Boldi et al. [3] anonymize a graph by anonymizing its edges, such that the utility of the graph in term of certain properties is maintained. Zhou and Pei [13] present an anonymization technique based on anonymizing graph neighborhoods. Liu and Terzi [9] present a technique by which every node has at least k other nodes with the same degree. Bonchi, Gionis, and Tassa [4] employ an information-theoretic methodology in perturbing edges in order to generate an anonymized graph. We use their anonymization technique in this study. To the best of our knowledge, we are the first to investigate the vulnerability of an anonymized feature matrix in terms of predicting node-overlap between the anonymized and auxiliary graphs.

The current literature is full of techniques which attempt to break privacy when the topology of the anonymized graph is released. Backstrom et al. [1] consider a scenario where an adversary creates a distinguishable subgraph before the graph is anonymized and released. They show that under certain conditions, the adversary can identify users in the anonymized graph using the structure that they earlier created. Using a set of seed matches, Narayanan and Shmatikov [11] show how the identities of the remaining nodes can be matched. Pedarsani and Grossglauser [12] study conditions under which the identities of nodes in an anonymized graph could be revealed. Korula and Lattanzi [8] assume that the anonymized graph and the auxiliary graphs have been generated from the same underlying graphs and break privacy of nodes by matching edges. Henderson et al. [7] propose a framework to recursively extract structural features of a node and use the structural features to match nodes across graphs. Gilpin, Eliassi-Rad, and Davidson [5] use guided role discovery to re-identify nodes. None of the previous work considers producing a guide for an adversary to select the most appropriate auxiliary data.

5. CONCLUSIONS

We considered the problem of predicting the amount of node-overlap between a known auxiliary graph and a feature matrix containing local structural properties of an anonymized graph. This problem is relevant to a variety of deanonymization tasks. Specifically, before an adversary can use an auxiliary graph (with known node-identities) for deanonymization, he/she must first select that auxiliary graph. Without further information, it makes sense for him/her to select the auxiliary graph whose nodes overlap with the anonymized data as much as possible, and so he/she needs to accurately predict the amount of node-overlap between the anonymized and auxiliary data.

We discussed challenges surrounding the aforementioned problem. In particular, a major challenge is that the structural characteristics of nodes that are present in both graphs versus those that are present only in the auxiliary graph may be very similar to one another. We presented examples showing degree distributions for both 'present' and 'absent' nodes. While there were some minor differences, these differences did not generalize to other graphs. Moreover, when applying a classifier (such as a support vector machine) to the task of differentiating between 'present' and 'absent' classes of nodes, it performed very poorly even when restricted to a single domain (such as academic coauthorships).

Despite the challenges, we showed that an adversary can make progress in estimating the amount of node-overlap between the anonymized and auxiliary data. We considered two possible cases. In the first case, the adversary has no information about the anonymized graph other than its structural feature matrix. In the second case, the adversary has obtained some supplemental information.

Without any supplemental (seed) information, an adversary can compare the centroids of the anonymized and auxiliary feature matrices, and weigh this value by the maximum possible node-overlap between the two graphs (which is based on their relative sizes). We find that this measure consistently overestimates the true node-overlap between the graphs. Thus, if this estimate is low, the adversary can be confident that the true node-overlap between the graphs is also low, and should thus attempt to find a different auxiliary graph for his/her deanonymization task.

If the adversary can somehow obtain a random unbiased sample of nodes from the anonymized graph that are labeled as 'present' or 'absent' in the auxiliary graph, he/she can trivially estimate the true node-overlap by simply measuring the ratio of 'present' to 'absent' nodes. However, it may be unrealistic to assume that the adversary can obtain an unbiased sample. More interestingly, if he/she can somehow obtain a sample of seed *matches* that are marked not only as 'present' or 'absent', but whose matching node in the auxiliary graph is known, then he/she can estimate the amount of structural change between nodes in the auxiliary and anonymized data. Notably, this sample of matches need not have the proper class balance between 'present' and 'absent' nodes; and it need not have any absent nodes at all. We observed that if the amount of structural change is low in this case, and the adversary has predicted a very high

node-overlap between the two graphs (i.e., their feature matrices have similar means), then his/her prediction is likely to be correct.

Moreover, we demonstrated that if an adversary is able to obtain 'present' or 'absent' labels for nodes in an auxiliary graph, describing whether those nodes are in the anonymized data, he/she can use this information to make informed predictions about a second auxiliary graph. In this way, he/she can predict the node-overlap between that second auxiliary graph and the anonymized data—effectively doing transfer learning.

Although the problem of predicting the most appropriate auxiliary graph for breaking privacy of an anonymized graph (released as a structural feature matrix) is extremely challenging, we have shown that an adversary can often make accurate predictions about the node-overlap between the anonymized and auxiliary data.

6. ACKNOWLEDGMENTS

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344; and supported in part by NSF Grant No. CNS-1314603, DTRA Grant No. HDTRA1-10-1-0120, and DAPRA under SMISC Program Agreement No. W911NF-12-C-0028.

The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. 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 DOE, LLNL, NSF, DTRA, DARPA, or the U.S. Government.

7. REFERENCES

[1] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural stenography. In *WWW*, pages 181–190, 2007.

[2] M. Berlingerio, D. Koutra, T. Eliassi-Rad, and C. Faloutsos. Network similarity via multiple social theories. In *ASONAM*, pages 1439–1440, 2013.

[3] P. Boldi, F. Bonchi, A. Gionis, and T. Tassa. Injecting uncertainty in graphs for identity obfuscation. *PVLDB*, 5(11):1376–1387, 2012.

[4] F. Bonchi, A. Gionis, and T. Tassa. Identity obfuscation in graphs through the information theoretic lens. *Inf. Sci.*, 275:232–256, 2014.

[5] S. Gilpin, T. Eliassi-Rad, and I. N. Davidson. Guided learning for role discovery (GLRD): Framework, algorithms, and applications. In *KDD*, pages 113–121, 2013.

[6] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis. Resisting structural re-identification in anonymized social networks. *PVLDB*, 1(1):102–114, 2008.

[7] K. Henderson, B. Gallagher, L. Li, L. Akoglu, T. Eliassi-Rad, H. Tong, and C. Faloutsos. It's who you know: Graph mining using recursive structural features. In *KDD*, pages 663–671, 2011.

[8] N. Korula and S. Lattanzi. An efficient reconciliation algorithm for social networks. *ArXiv e-prints*, July 2013.

[9] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *SIGMOD*, pages 93–106, 2008.

[10] A. Narayanan and E. W. Felten. No silver bullet: De-identification still doesn't work. Policy report, Princeton University, Princeton, NJ, July 2014. Available at `http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf`.

[11] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Symposium on Security and Privacy*, pages 173–187, 2009.

[12] P. Pedarsani and M. Grossglauser. On the privacy of anonymized networks. In *KDD*, pages 1235–1243, 2011.

[13] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *ICDE*, pages 506–515, 2008.