De-anonymizing Social Graphs via Node Similarity

Hao Fu
University of Science and Technology of China
fuch@mail.ustc.edu.cn

Aston Zhang
University of Illinois at Urbana-Champaign
lzhang74@illinois.edu

Xing Xie
Microsoft Research
xingx@microsoft.com

ABSTRACT

Recently, a number of anonymization algorithms have been developed to protect the privacy of social graph data. However, in order to satisfy higher level of privacy requirements, it is sometimes impossible to maintain sufficient utility. Is it really easy to de-anonymize "lightly" anonymized social graphs? Here "light" anonymization algorithms stand for those algorithms that maintain higher data utility. To answer this question, we proposed a de-anonymization algorithm based on a node similarity measurement. Using the proposed algorithm, we evaluated the privacy risk of several "light" anonymization algorithms on real datasets.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration—*Security, integrity, and protection*; K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

Keywords

De-anonymization; privacy protection; social network

1. INTRODUCTION

In social networking sites, users and their social ties can be described as a *social graph*. In order to satisfy the need for analysis, operators of social networking services are increasingly sharing information that could potentially breach users' privacy. An adversary could leverage auxiliary information such as node degrees, neighborhoods of breached nodes, or subgraphs of arbitrary sizes nearby certain nodes, to reveal the identities of nodes in the real world.

To preserve privacy, the data need to be *anonymized* before publishing. Anonymization is performed by modifying the structure and descriptive information of social graphs. A straightforward approach, namely the naïve anonymization, removes personal identifiable information such as names and leaves the graph structure as it was. Studies have shown the privacy risk of this approach [1, 3], and a number of anonymization algorithms were proposed [2, 4]. Modifying the graph increases the difficulty of attacks but compromises the utility of data at the same time, rendering the anonymized graphs less useful for analysis. As it is sometimes intractable to collect auxiliary information and social graphs are evolving over time, naïve anonymization or "light" anonymization algorithms that only make minor modifications might work much better than expected in practice.

This leads to an interesting question: is it really easy to de-anonymize those social graphs? In this paper, we propose a graph node similarity measurement in consideration with both graph structure and descriptive information, and a de-anonymization algorithm based on the measurement. As it is infeasible to define precisely what "light" algorithms are due to the different behaviors of the algorithms, we choose several typical anonymization algorithms as their representatives to evaluated with. Our results showed that the proposed algorithm was efficient and effective to de-anonymize social graphs without any initial *seed mappings* [5].

2. DE-ANONYMIZING

A *simple graph* is an undirected graph $G = (V, E)$ without any descriptive information. Nodes in V correspond to users, and an edge $(i, j) \in E$ indicates a social tie between users i and j. A *rich graph* is the combination of a directed or undirected graph, and two attribute sets X and Y. User i's descriptive information (or *node-attribute*) is denoted as $X(i)$, and the description of a social tie (i, j) (or *edge-attribute*) is represented as $Y(i, j)$.

We refer to the anonymized and published social graph as the *target graph*. We assume that an adversary can always collect a subgraph, namely *auxiliary graph*, nearby nodes of interest. This assumption is practical because online social networking sites are usually partially or fully accessible. Note that this is the only prior knowledge: The adversary does not know the real identity, or seed mapping, of any node in the target graph. The goal of de-anonymization is to find *identity disclosures* in the form of one-to-one mappings as many and accurate as possible.

Simple graph Suppose we are trying to compare the auxiliary graph $G_1 = (V_1, E_1)$ and the target graph $G_2 = (V_2, E_2)$. For nodes $i \in V_1$ and $j \in V_2$, we introduce the *node similarity score* $\mathcal{S}(i, j)$ as a structural measurement on how similar the two nodes are. The node similarity score is defined recursively. First, the initial values are taken as $\mathcal{S}^{(0)}(i, j) = 1$. Denoting i's neighbor nodes as $N_1(i)$ and j's neighbor nodes as $N_2(j)$, we construct a complete bipartite graph $B_{i,j} = (N_1(i), N_2(j), N_1(i) \times N_2(j))$ by weighting edge

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577366.

(i', j') as $\mathcal{S}^{(k)}(i', j')$. We then find the maximum weighted matching $\theta_{i,j}$ of $B_{i,j}$, where node $l \in N_1(i)$ is matched to node $\theta_{i,j}(l) \in N_2(j)$. Finally, $\mathcal{S}^{(k+1)}(i, j)$ is assigned as

$$\mathcal{S}^{(k+1)}(i, j) = \sum_{l \in N_1(i)} \mathcal{S}^{(k)}(l, \theta_{i,j}(l)) \quad (1)$$

Here the match $\theta_{i,j}$ is re-calculated in every iteration with the node similarity scores $\mathcal{S}^{(k)}$. The calculation is repeated until the normalized scores converge, and the normalization is done by dividing $\mathcal{S}^{(k)}$ by the maximum $\mathcal{S}^{(k)}(i, j)$.

We again construct a complete bipartite graph $B = (V_1, V_2, V_1 \times V_2)$ by weighting (i, j) as $\mathcal{S}(i, j)$, and the identity disclosures are found as the maximum weighted match of B. As node pairs with higher similarity scores are more likely to be correct, only the top M (which is specified by the adversary) mappings are outputted. Note that methods to produce identity disclosure are not limited to the proposed approach. For example, a ranking of candidates for each node can be produced by sorting the candidates' similarity scores. The adversary can later check top candidates manually by comparing the profiles with domain knowledge.

Rich graph The *node-attribute similarity* $\mathcal{S}_X(i, j)$ represents similarity between node-attribute sets $X(i)$ and $X(j)$. Analogously, the *edge-attribute similarity* $\mathcal{S}_Y(i_1, j_1, i_2, j_2)$ measures how similar two edge-attribute sets $Y(i_1, j_1)$ and $Y(i_2, j_2)$ are. In directed graphs, there could be two edges of opposite directions between two nodes. The *relation similarity* $\mathcal{S}_R(i_1, j_1, i_2, j_2)$ measures the similarity of node pairs (i_1, j_1) and (i_2, j_2) in conjunction with edges of both directions. We only assume the measurements range from 0 (completely different) to 1 (possible equivalent) inclusively. The node similarity defined in Equation (1) is extended as

$$\begin{aligned} \mathcal{S}^{(k+1)}(i, j) &= \alpha \cdot \sum_{l \in N_1(i)} \mathcal{S}^{(k)}(l, \theta_{i,j}(l)) \cdot \mathcal{S}_R(i, l, j, \theta_{i,j}(l)) \\ &+ \mathcal{S}_X(i, j) \end{aligned} \quad (2)$$

The constant factor α trades off the importance of node-attribute against graph structure and edge-attribute. The initial values are taken as $\mathcal{S}^{(0)}(i, j) = \mathcal{S}_X(i, j)$.

3. EXPERIMENTS

Simple graph We used a co-author graph from Microsoft Academic Search that was published in WSDM 2013 Data Challenge for our evaluation. It consists of 8,248 nodes and 18,732 edges without any attribute. Every node corresponds to an author, and two authors are linked by a single edge only if they have collaborated at least one paper.

Rich Graph We extracted a social graph of 2.3 million nodes and 55.4 million directed edges from the Tencent Weibo dataset that was published in KDD Cup 2012. The node attribute set contains gender, birth year, and the number of user tweets, and the edge attribute describes the following relationship and how many times a user has mentioned, retweeted, or left a comment for another user.

In our experiments, we chose naïve anonymization, k-degree anonymity (not applicable for rich graphs), and randomization to anonymize target graphs. We implemented two algorithms proposed by Liu and Terzi [4] for k-degree anonymity: the one that only adds edges, and the one that adds and deletes edges simultaneously. Three different approaches for randomization were evaluated: sparsification, perturbation, and switching [2]. The test data were generated from the original graph by extracting 10 pairs of subgraphs (the auxiliary graph G_1 and the target graph G_2) randomly with specified overlap ($\beta_V = |V_1 \cap V_2|/|V_1 \cup V_2|$). Copies of G_2 were anonymized with different algorithms. For the rich graph, G_1 is taken as a subgraph of 10,000 nodes from the original graph, and the target graph is induced from a part of the auxiliary graph together with the remaining 2.3 million nodes. We applied our algorithm on the graph pairs and the result is reported as the averages.

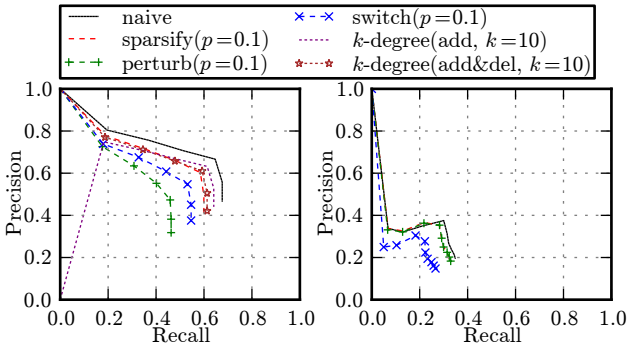

Figure 1: Accuracy of attacks with different M (the number of mappings) on a simple graph (left, 50% overlap) and a rich graph (right, 5,000 overlap)

The result (Figure 1) showed that the precision and recall of our attacks were reasonably high. The relatively low performance of attacks on rich graphs was unsurprising, since the node overlap of rich graph pairs is only 0.22%. The result also suggested that the attributes played a more important role in de-anonymizing rich graphs, since altering the graph structure did not make much difference.

Acknowledgements We would like to gratefully acknowledge the organizers of WSDM 2013 Data Challenge and KDD Cup 2012, as well as Microsoft Academic Search and Tencent Inc. for providing the datasets.

4. REFERENCES

[1] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*, pages 181–190, 2007.

[2] F. Bonchi, A. Gionis, and T. Tassa. Identity obfuscation in graphs through the information theoretic lens. In *Proceedings of 27th International Conference on Data Engineering (ICDE 2011)*, pages 924 –935, 2011.

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

[4] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *Proceedings of the 2008 ACM International Conference on Management of Data (SIGMOD 2008)*, pages 93–106, New York, NY, USA.

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