Privacy In Dynamic Social Networks

Smriti Bhagat
Rutgers University
smbhagat@cs.rutgers.edu

Graham Cormode
AT&T Labs–Research
graham@research.att.com

Balachander Krishnamurthy
AT&T Labs–Research
bala@research.att.com

Divesh Srivastava
AT&T Labs–Research
divesh@research.att.com

ABSTRACT

Anonymization of social networks before they are published or shared has become an important research question. Recent work on anonymizing social networks has looked at privacy preserving techniques for publishing a single instance of the network. However, social networks evolve and a single instance is inadequate for analyzing the evolution of the social network or for performing any longitudinal data analysis. We study the problem of repeatedly publishing social network data as the network evolves, while preserving privacy of users. Publishing multiple instances of the same network independently has privacy risks, since stitching the information together may allow an adversary to identify users in the networks.

We propose methods to anonymize a dynamic network such that the privacy of users is preserved when new nodes and edges are added to the published network. These methods make use of link prediction algorithms to model the evolution of the social network. Using this predicted graph to perform group-based anonymization, the loss in privacy caused by new edges can be reduced. We evaluate the privacy loss on publishing multiple social network instances using our methods.

1. INTRODUCTION

Social networks are now a ubiquitous feature of modern life. Facebook claims over 300 million accounts, while Twitter has long-term plans for a billion users. As a result, the study of social networks and their associated ecosystem (applications, third party tools, cross-network interactions) has become a core topic in the analysis and measurement communities. A key feature of the current generation of social networks, exemplified by Facebook, is that the full picture is not visible to the outside observer. The network is defined by the demographic information of the users, along with the detailed connection information: not just "friend" links, but the richer set of messages, public postings, application uses and so on. This detailed information of a user is not visible unless that user has given explicit permission to the viewer. So currently only the social network operator has the full view, leaving other interested parties—network researchers, sociologists, application designers and other players in the social space—to scrape away at the edges.

A full release of snapshots of the network would address this need. However, the default settings are private for a reason: social networks contain much sensitive personal information entered by their users. Users have an expectation that their information is only visible to their friends and are very vocal in their opposition when these expectations are not met. This has been demonstrated by the opposition to Facebook's Beacon advertising platform and by the regular news stories about firings and scandals resulting from unintended information disclosure[1].

Instead, the computer science community has proposed *anonymization* of social network data as a mechanism for letting interested third-parties see the structure of social networks, without revealing private information. This is an evolving field of research: methods are proposed, and analyzed for weaknesses. Simplistic anonymization methods have been shown to be vulnerable to attack by an adversary with a lot of information about the network, or the ability to inject structures into the network [1, 7]. In response, more sophisticated anonymizations have been proposed and evaluated [9].

However, a limitation of the work so far in this area is that the focus has primarily been on *static* networks. That is, the dataset is considered to be a single instance of the network, represented as a graph. This fails to capture the highly dynamic nature of social network data. Recently, significant effort has been made to analyze the temporal evolution of social networks [4, 6, 8], so it follows that any attempt at anonymization should also describe the time-varying nature of the data. In particular, our model should not be of a single anonymized snapshot of the network being released, but rather updates reflecting the current state of the network released on a regular basis: every few months, say, for timely analysis.

For the dynamic case, it is more challenging to ensure the necessary levels of privacy while still keeping the output relevant for its intended uses. We argue that it is not sufficient to anonymize each version of the network independently: this is easily shown to leak information by comparing the different versions of the data. Instead, we maintain that it is necessary for subsequent releases to be consistent with the initial release. Consequently, the decisions made for an initial anonymization affect later iterations. Since these decisions are fixed, the subsequent arrival of new information has implications for future releases: the subsequent release

WWW2010, April 26-30, 2010, Raleigh, North Carolina.
.

[1] See http://gigaom.com/2007/11/06/facebook-beacon-privacy-issues/, http://w2.eff.org/Privacy/AOL/

may increase the amount of information (measured in terms of probabilities) that can be extracted about the behavior of individuals in the data. Unfortunate decisions early on mean that later releases lead to higher probabilities than desired, and may require that some information is suppressed from the subsequent releases. It may even be necessary to halt the release process completely, so that no private information is leaked. The question then arises, how to choose the anonymizations early on, so that the increase in probabilities for later releases is minimized, without knowing in advance how the network will grow?

2. PROBLEM STATEMENT

Let $\mathcal{G} = \{G_1, G_2, \ldots, G_T\}$ be a sequence of T graphs representing the network observed at timesteps $t = 1, 2, \ldots, T$ respectively. Here, $G_t = (V_t, E_t)$ is a graph representation of a time-varying social network, such that V_t is the set of vertices that represent users (or, entities) U_t that are a part of the network at time t, and E_t is the set of all edges (interactions between users) created up to time t. Given $\mathcal{G}$ as input, our objective at any time t is to publish an anonymized version of graph G_t as G'_t (without knowledge of any future G_i for $i > t$), so that the anonymization satisfies defined privacy and utility goals.

We outline a solution, which leverages the existence of *link prediction algorithms*. That is, we are able to use the current state of the network to predict structures that are more likely to arise in future steps. This information is then used in conjunction with the current state to choose an anonymization which is currently safe, and which is expected to remain safe and useful for future releases. Implementing this concept requires us to specify many details. We must identify a set of possible prediction methods that are appropriate for this task. We do not treat the predicted edges equally to observed edges, and must define how to incorporate them into the anonymization algorithms. By following this outline we arrive at methods which can effectively anonymize dynamic social network data with only small increases in the probabilities.

More precisely, our solution extends the schemes for anonymizing a graph proposed in [2]. There, the anonymization is provided by masking the mapping between nodes in the graph V_t and entities U_t such that each $v \in V_t$ is associated with a *list* of possible labels $l(v) \subset U_t$ such that $v \in l(v)$. The underlying graph structure is published, with a label list at each node instead of the user identifier. The lists can therefore be generated by partitioning the nodes into *groups*, so that each node in the group is given the same list, which consists of all the (true) labels of nodes in the group. The core of our approach is based on the idea of leveraging *link prediction*. The problem of link prediction has been heavily studied in the link analysis and mining literature [5]. Here, we use link prediction to predict which links are likely to arise in the future. The motivation for predicting the network evolution before grouping is to preempt the weakening of privacy resulting from the addition of edges to the published graph. If the model predicts a significant fraction of the edges that appear in the future, the privacy guarantees are expected to remain intact as the graph evolves. It is also necessary to select a subset of the predicted edges to be able to find a usable anonymization, so we explore a variety of methods to prioritize these predicted edges. Based on the predicted edges we choose how to group the nodes in such a way that not only do existing edges meet an appropriate grouping condition, but also future edges are unlikely to violate it either.

3. RESULTS

We omit the full details of our methods and their evaluation from this brief summary, and refer the interested user to our full report [3]. The main contributions of this work are as follows:

- We define the problem of anonymization of dynamic graphs, and describe requirements of the output.
- We describe our approach to the problem by choosing an initial anonymization based on prediction which is expected to be more resilient to future evolution of the network. We provide metrics for evaluating the privacy preserving quality of an anonymization.
- We explain how different prediction models can be incorporated into our framework, and how the results of the prediction can be fine-tuned to help create our anonymization by adoption of appropriate conditions.
- We perform a set of experiments over temporal data representing social network activity and empirically evaluate the privacy guarantees and utility resulting from the anonymization methods we propose. Our study shows that with the correct choice of prediction method and anonymization properties, it is possible to provide useful data on dynamic social networks while retaining sufficient privacy.

4. REFERENCES

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

[2] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava. Class-based graph anonymization for social network data. In *VLDB*, 2009.

[3] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava. Prediction provides privacy in dynamic social networks, 2010.

[4] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *ACM SIGKDD*, 2008.

[5] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.

[6] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the flickr social network. In *ACM Workshop on Online Social Networks*, 2008.

[7] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Symposium on Security and Privacy*, 2009.

[8] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *ACM Workshop on Online Social Networks*, 2009.

[9] B. Zhou, J. Pei, and W. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explor. Newsl.*, 2008.