Automatic Sanitization of Social Network Data to Prevent Inference Attacks

Raymond Heatherly
The University of Texas at Dallas
Department of Computer Science
Richardson, Texas, USA
rdh061000@utdallas.edu

Murat Kantarcioglu
The University of Texas at Dallas
Department of Computer Science
Richardson, Texas, USA
muratk@utdallas.edu

ABSTRACT

As the privacy concerns related to information release in social networks become a more mainstream concern, data owners will need to utilize a variety of privacy-preserving methods of examining this data. Here, we propose a method of data generalization that applies to social networks and present some initial findings for the utility/privacy tradeoff required for its use.

Categories and Subject Descriptors

K.4.1 [**Public Policy Issues**]: Social Network Privacy

General Terms

Experimentation, Measurement, Security

Keywords

Social Network privacy, sanitization

1. INTRODUCTION

It is generally accepted that social networks are both growing larger and becoming more important. The data contained in these diverse networks is useful for many reasons. Advertisers want access to this data so that they can know their audience better and thus target ads more effectively. Governments may want access to be able to track the habits and behaviors of its citizens or of potential terror threats.

However, privacy concerns prevent many data owners from being able to release at least some of the data that they hold. When we deal with attacks on privacy in social networks, we generally consider two threats. The first is an identification attack. This is a situation where an attacker attempts to determine the real-world identity from examining a social network's data. For instance, in [1], the authors use various techniques to anonymize the link structure of the social network to prevent identification attacks, assuming that the links are the extent of the network released.

Second is an inference attack. This method of attack assumes that there is some hidden data within the social network's data. The attacker's goal is to use various machine learning techniques in an attempt to predict this hidden data. In other words, an attacker tries to build a highly accurate classifier to predict hidden sensitive data. In [3], the authors use inference attacks based upon the link structure to recommend hiding from release any data that they identify as the highest threat. In [2], authors provide intelligent removal techniques that can remove both links and details to prevent inference attacks.

In this paper, we focus solely on the inference problem. Here, we attempt to solve the problem of inference attacks on social network privacy through a novel generalization approach to network sanitization. Additionally, we define a new form of data generalization method, which we refer to as detail value decomposition. We believe that our contributions improve the area of social network anonymization. Mainly, all data contained within the graph is accurate. That is, we do not generate false data to replace details with similar, yet unrelated traits. Additionally, unlike other anonymization techniques, we do not alter the edge set of the graph in any way. The second, related, benefit is that our model specifically acknowledges a utility/privacy tradeoff. Since there is no option of perfect privacy with a data release, we believe that this corresponds to a valid real-world data release.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

2. PRIVACY DEFINITION

We define the following model of privacy for social networks:

DEFINITION 1. *A Social Network is represented as a graph, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{D}\}$, where $\mathcal{V}$ is the set of nodes in the graph, $\mathcal{E}$ is the set of edges connecting those nodes, and $\mathcal{D}$ is the set of details containing some personal information about the members of $\mathcal{N}$. Here an individual detail is a fact that a person releases about himself, such as "N_i's Favorite Book is 'Harry Potter"'.*

DEFINITION 2. *Background knowledge, $\mathcal{K}$, is some data that is not necessarily related to the social network, but that can be obtained through various means by an attacker. Examples of background knowledge include voter registration, election results, phone book results, etc.*

DEFINITION 3. *A graph is $(\Delta, \mathcal{C}, \mathcal{G}, \mathcal{K})$-private if, for a given set of classifiers $\mathcal{C}$,*

$$\max\left[\left(\max_{c\in\mathcal{C}}\left(\mathcal{P}_{c(\mathcal{G},\mathcal{K})}\right) - \max_{c'\in\mathcal{C}}\left(\mathcal{P}_{c'(\mathcal{K})}\right)\right), 0\right] = \Delta$$

That is, if we have any set of given classifiers, $\mathcal{C}$, then the classification accuracy of any arbitrary classifier $c' \in \mathcal{C}$ when trained on $\mathcal{K}$ and used to classify $\mathcal{G}$ to predict sensitive hidden data is denoted by $\mathcal{P}_{c'(\mathcal{K})}$. Similarly, $\mathcal{P}_{c(\mathcal{G},\mathcal{K})}$ denotes the

prediction accuracy of the classifier that is trained on both $\mathcal{G}$ and $\mathcal{K}$. Here Δ denotes the additional accuracy gained by the attacker using $\mathcal{G}$. Ideally, if $\Delta = 0$, this means that the attacker does not gain additional accuracy in predicting sensitive hidden data.

3. GENERALIZATION

In order to combat inferrence attacks on privacy, we attempt to provide detail anonymization for social networks. By doing this, we believe that we will be able to reduce the value of $\Delta_{\mathcal{G},\mathcal{K}}(C)$ to an acceptable threshhold value that matches the desired utility/privacy tradeoff for a release of data.

DEFINITION 4. *A detail generalization hierarchy (DGH) is an anonymization technique that generates a hierarchical ordering of the details expressed within a given category. The resulting hierarchy is structured as a tree, but the generalization scheme guarantees that all values substituted will be an ancestor, and thus at a maximum may be only as specific as the detail the user initially defined.*

To clarify, this means that if a user inputs a favorite activity as the Boston Celtics, we could have, as an example, the following DGH: Boston Celtics → NBA → Basketball. This means that to completely anonymize the entry of "Boston Celtics" in a user's details, we replace it with "Basketball." However, notice that we also have the option of maintaining a bit more specificity (and therefore utility) by replacing it instead with "NBA." This heirarchical nature will allow us to programmatically determine a more efficient release anonymization, which hopefully ensures that we have a generalized network has more information than simply deleting details. Our scheme's guarantee, however, ensures that at no time will the value "Boston Celtics" be replaced with the value "Los Angeles Lakers."

Alternately, we have some details, such as "Favorite Music" which do not easily allow themselves to be placed in a hierarchy. Instead, we perform detail value decomposition on these details.

DEFINITION 5. *Detail Value Decomposition (DVD) is a process by which an attribute is divided into a series of representative tags. These tags do not necessarily reassemble into a unique match to the original attribute.*

Thus, we can decompose a group such as "Enya" into {ambient, alternative, irish, new age, celtic} to describe the group.

To generate the DGH and the DVD, we use subject authorities which, through having large amounts of data, are able to provide either a DGH or a DVD for some particular (set of) detail types.

4. EXPERIMENTS

The data used in these experiments was pulled from the Dallas/Fort Worth network of Facebook in April of 2007. That is, individuals who joined Facebook identified themselves as being a part of a then-regional network. This data set is comprised of approximately 167,000 nodes with 3 million links and 4.5 million details listed.

Each of those details fell into one of several categories: Religion, Political Affiliation, Activities, Books, Music, Quotations, Shows/Movies, and Groups. Due to the lack of a reliable subject authority, Quotations were discarded from all experiments. To generate the DGH for each Activity, Book, and Show/Movie, we used Google Directories. To generate the DVD for Music, we used the Last.fm tagging system. To generate the hierarchy for Groups, we used the classification criteria from the Facebook page of that group.

To account for the free-form tagging that Last.fm allows, we also store the popularity for each tag that a particular detail has. This font size is representative of how many users across the system have defined that particular tag for the music type. We then keep a list of tag recurrence (weighted by strength) for each user. For Music anonymization, we eliminate the lowest-scoring tags.

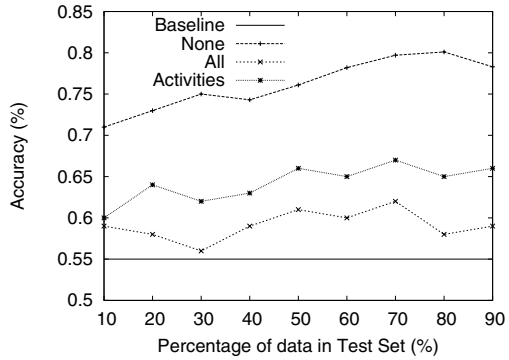

Figure 1: Effect of generalization on utility

In our experiments, we assume that the trait "political affiliation" is a sensitive attribute that the data owner prefers to hide. Our $\mathcal{C}$ include a naive Bayes classifier and the implementation of SVM from Weka.

In Figure 1, we present some initial findings from our domain generalization. We present a comparison of simply using $\mathcal{K}$ to guess the most populated class from background knowledge, the result of generalizing all trait types, generalizing no trait types, and when we generalize the best single performing trait type (activities).

We see here that our method of generalization (seen through the All and Activities lines) do indeed decrease the accuracy of classification on the data set. Additionally, our ability to classify on **non-sensitive attributes (such as gender) are affected by only 2-3%** over the course of the experiments (figure omitted for space). Interestingly, while previous work[2] indicates that Group membership is the dominant detail in classification, we see the most benefit here from generalizing only the Activities detail. For full comparison and discussion, please see the full paper.

5. ACKNOWLEDGMENTS

This work was partially supported by Air Force Office of Scientific Research MURI Grant FA9550-08-1-0265, National Institutes of Health Grant 1R01LM009989, National Science Foundation (NSF) Grant Career-0845803, and NSF Grant CNS-0964350,CNS-1016343

6. REFERENCES

[1] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. *University of Massachusetts Technical Report*, pages 07–19, 2007.

[2] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *WWW Poster*, 2009.

[3] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540. ACM, 2009.