Measuring topological anonymity in social networks

Lisa Singh
Georgetown University
Computer Science Department
Washington, DC 20057
singh@cs.georgetown.edu

Justin Zhan
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
justinzh@andrew.cmu.edu

Abstract

While privacy preservation of data mining approaches has been an important topic for a number of years, privacy of social network data is a relatively new area of interest. Previous research has shown that anonymization alone may not be sufficient for hiding identity information on certain real world data sets. In this paper, we focus on understanding the impact of network topology and node substructure on the level of anonymity present in the network. We present a new measure, topological anonymity, that quantifies the amount of privacy preserved in different topological structures. The measure uses a combination of known social network metrics and attempts to identify when node and edge inference SMPCes arise in these graphs.

1 Introduction

Social network analysis has emerged as a key analysis technique for sociologists, anthropologists, biologists and economists. Typically when we consider social network data, we view it as data that is available to the public. However, many social networks are now being automatically extracted from private data sources. Examples include social networks derived from corporate email servers, customer referral databases, personal medical records, and disease population databases. Assuming only a simple network with a single node type and a single edge type, we are interested in knowing the level of privacy preserved for different network topologies. Are nodes obscured more in a network containing a large number of triangles or stars? Are relationships between nodes more apparent when local neighborhoods have certain topological structures? Can we use the topological structure to measure the level of anonymity in the network? Finally, what measures are reasonable for quantifying privacy in different topologies? To study some of the behaviors associated with social networks, how accurate do the network measures need to be for data mining applications, e.g. clustering, community discovery, prominent node identification, etc.? While we anticipate many of these topics will be explored over the next few years, the goal of this paper is to begin the conversation by analyzing known network topologies in the context of privacy preservation and proposing a measure for quantifying the privacy level of the network. This measure can also be used to evaluate the effectiveness of different perturbation strategies, i.e. does the removal of a particular edge decrease the level of anonymity for a particular group of nodes?

This paper is organized as follows. The next section overviews some of the related literature. We then define privacy breaches for both nodes and edges in section 3 and discuss graph topology measures in that context. In section 4 we introduce our topological anonymity measure and describe it using a simple example. Section 5 presents experiments where we evaluate topological anonymity on two synthetic data sets and one real world data set. Finally, conclusions and future directions are presented in section 6.

2 Related Literature

Very little has been written on privacy preservation of social network data. Research in this area began using an interactive approach, where the user submits a question to a database and receives a noisy version of the actual result [4]. Two recent papers focusing on non-interactive privacy preservation of social networks investigate different types of privacy breaches [3, 6]. Backstrom et. al. show that naive anonymization (replacing actual values of identifying attributes with synthetic ones) of both passive and active attacks can lead to significant breaches, where an active attack involves a malicious adversary who ’plants’ nodes into the network prior to anonymization. Hay et. al. also show that naive anonymization does not always prevent inference related privacy breaches [6]. In that paper, they use random perturbation to delete and insert edges. While these papers discuss the general impact of topology on privacy, they do not analyze the ’obscuring’ ability of different topological structures. By understanding this relationship, users can gain insight into the level of privacy anonymity alone can provide for social networks.

K-anonymity was introduced for privacy preservation of independent, unlinked data records. Each individual should not be distinguishable from $k - 1$ other individuals [7]. We mention this line of work since others have described network privacy in terms of k-anonymity. However, because our nodes are not independent and are linked together, we believe k-anonymity as identified in [7] is difficult to achieve in graphs where dependencies exist in the data. Two nodes that are indistinguishable across some node structural metrics, does not guarantee that they are across other ones, particularly path related measures across the network. However, if we limit anonymity to local neighborhood structure of a node, k-anonymity can be an important approximation.

3 Measuring privacy breaches in social networks

In this paper, we consider a basic uni-mode social network G containing a set of nodes V of a single type and a set of edges E of a single type. Formally, $G = (V, E)$, where $V = \{v_1, v_2 ... v_n\}$ and $E = \{(v_i, v_j) \,|\, v_{i,} v_j \in V, i \neq j, 1 \leq i, j \leq n\}$. Examples include people connected by friendship, email, co-membership on a team, or colleagues at work.

3.1 Local neighborhood privacy breaches

We begin by describing two local neighborhood privacy breaches, a node identity breach and an edge inference breach. Since we are interested in breaches based on topological structure, our focus is on passive adversaries that do not have access to the data before it is released. The adversary knows who he is connected to, but is not certain if his neighbors know each other.

A **node identity breach** occurs if the label of a known network participant is determined after the graph has been anonymized and privatized. In Figure 1, if an adversary can label any of the nodes with complete certainty, the node position in the network is then known and a node identity breach has occurred.

Suppose that participant v_x is determined to be one of p possible actors. A **partial node identity breach** occurs if $p < \epsilon$, where ϵ is a threshold for the required level of anonymity for every node in the network. Here, the level of anonymity refers to the position of the node in the graph. If a node has a unique structural location in the graph, then the likelihood of labelling the node increases. For our example in Figure 1, if a connection exists between nodes B and C, nodes A and D are indistinguishable after anonymization. They both maintain the same position in the graph. Therefore, if the adversary is searching for node A, i.e. a node connected to two other nodes, he can narrow it down to one of two nodes based on the graph structure. If $\epsilon = 2$, a node identity breach does not occur. If ϵ is greater than two, a partial node identity breach has occurred. Determining the threshold or the level of anonymity that is reasonable is dependent on the needs of those releasing the data set.

An **edge inference breach** occurs if an adversary is able to determine whether or not two of its neighbors have an edge between them. For example, node A is connected to nodes B and C in Figure 1. An edge inference breach occurs if A determines whether or not an edge exists between nodes B and C.

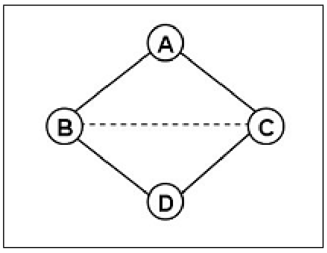

Figure 1. Social network example 1

3.2 Using social network metrics to indicate privacy

There are a number of measures we can use to locate node and edge position in the network based on graph theory [5]. Sociologists have also proposed tangential metrics that can be used to characterize and understand social networks. A common set of measures used for social network analysis are node centrality and neighborhood measures. For a detailed discussion of these measures, we refer you to [8]. In this paper, we will use a combination of two node and edge position metrics as the basis for evaluating the 'hiding ability' of a network's topology - degree and clustering coefficient:

- Degree $\deg(v_i)$ - The number of nodes directly connected to v_i. A node with high degree represents a well-connected individual in the network, i.e., one that has many direct relationships with other individuals. In Figure 1, $deg(A) = 2$.

- Clustering coefficient $CC(v_i)$ - The clustering coefficient of node v_i is a measure of the likelihood that v_i's neighbors know each other. It is the fraction of edges that exists between v_i's immediate neighbors divided by the possible number of edges that could exist [9]. The higher the clustering coefficient of a node, the more densely connected, e.g. 'clique-like', the local neighborhood of the node. In Figure 1, $CC(A) = 1$ if an edge exists between nodes B and C. If no edge exists, $CC(A) = 0$.

We will use these two network metrics to determine the 'position' of each node in the graph. We say that a node v_i is *hidden* if there are at least ϵ other nodes in the network with the same degree, $deg(v_i)$. However, even if a node is hidden, an edge inference breach may occur. We say that an edge inference is *avoided* if nodes with the same degree have different immediate neighborhood structures, i.e. clustering coefficients.

In Figure 2, nodes A and D both have a degree of 2. Because the degree is greater than one, some node anonymity exists. Suppose the adversary is node A and he is trying to determine if B and C know each other. In an unlabelled graph, the adversary may have one of two positions (A's or D's) in Figure 2. Because the neighbors have an edge between them in one case (node B is connected to node C) and not the other (node C is not connected to node E), the adversary A cannot be certain if nodes B and C are connected. A network that has a topology containing this type of node overlap and edge structure diversity throughout is more private than one that does not. Using this substructure information, we now propose a measure that quantifies the amount of anonymity in the structure of the network.

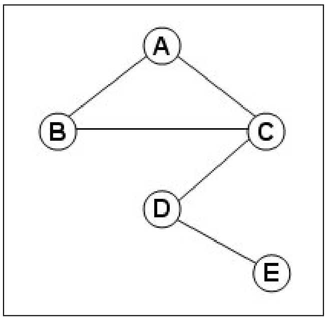

Figure 2. Social network example 2

4 Topology anonymity

When defining the level of anonymity associated with a social network, we consider variations of degree and clustering coefficient. We define D_a to be the set of nodes with degree a.

$$D_a = \{v_i \mid deg(v_i) = a \,\forall\, v_i \in V\}$$

This *degree set* contains all the nodes with a distinct degree values in the network. In Figure 2, $D_1 = \{E\}$, $D_2 = \{A, B, D\}$, and $D_3 = \{C\}$. $|D_a|$ represents the number of nodes in the set. For our example, $|D_1| = 1$, $|D_2| = 3$, and $|D_3| = 1$. As the number of nodes in D_a increases, node anonymity also increases. We use D to represent the list of degree sets in the network, $D = \{D_1, D_2, D_3, ...D_{max(deg(G))}\}$. Here, $D_{max(deg(G))}$ is the set of nodes with the maximum degree in the network G.

To compare the local neighborhood of nodes with the same degree, we introduce a boolean measure, CC_dif that assigns a value based on the variance in the clustering coefficients of nodes in a degree set.

$$CC\_dif_a = \begin{matrix} 0 & if \;\; var(CC(D_a)) = 0 \\ 1 & if \;\; var(CC(D_a)) > 0 \end{matrix}$$

If the variance var of the clustering coefficients for nodes of degree a in degree set D_a is zero, then CC_dif_a is zero. Otherwise, it is one. To keep our measure simple, we do not consider the level of the variance in the degree set. One could change the threshold, e.g., $CC\_dif_a = 0$ if $var(CC(D_a)) < \delta$ where δ represents a small, unallowable variance. While this may be useful, we leave that for future work. If there is only one node in D_a, the variance is undefined and cannot be calculated. In those cases, the node and edges in question are not hidden from the adversary and $CC\_dif_a = 0$. For the network in Figure 2, $CC\_dif_1 = 0$, $CC\_dif_2 = 1$, and $CC\_dif_3 = 0$.

We now integrate the degree set node information and the local neighborhood variance of a degree set into a single network privacy measure that enables the user to determine whether or not the topological structure of the network sufficiently obscures local node and edge positions.

Definition 1 ***Topological anonymity*** *(ta) represents the level of obscurity in the structure of a connected network where the number of nodes is at least three, $n \geq 3$.*

$$ta = \frac{\sum_{i=1}^{max(deg(G))}(|D_i| \times CC\_dif_i) - \sum_{j=1}^{\epsilon-1}|D_j|}{n}$$

The first component of the measure sums the product of the degree set and the boolean clustering coefficient variance. Then all the degree sets containing fewer than ϵ members are subtracted. Finally, to normalize the value, we divide by the number of nodes in the network. A topological anonymity value close to -1 is associated with a network having a large number of nodes in degree sets less than ϵ. As the value of ta increases, the hiding ability of the network structure increases. A topological anonymity value of 1 indicates a network that does not contain basic node identity breaches or edge inference breaches.

In order to simplify the metric, we assume that the graph is connected and that there are at least three nodes in the network. Extensions for graphs that contain multiple disconnected components is straightforward, but outside the scope of this paper.

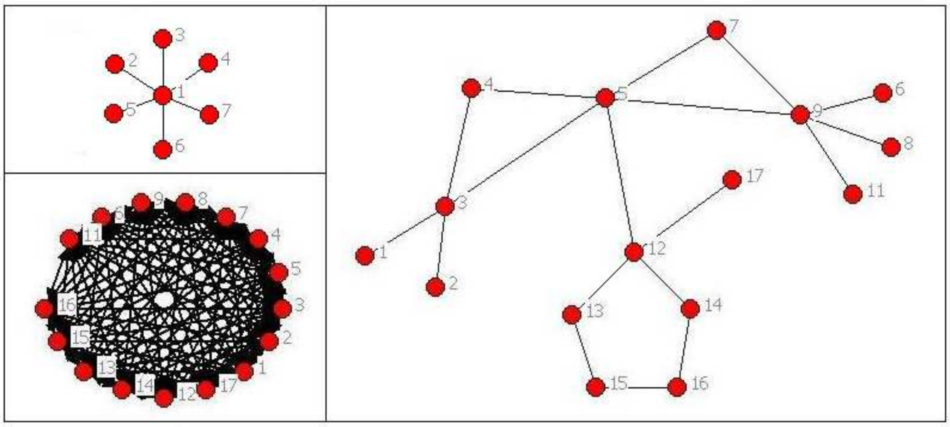

Figure 3. Varying values for ta - Upper left: $ta = -5/6$**; Lower left:** $ta = 0$**; Right:** $ta = 0.62$

Figure 3 shows three different networks and their corresponding ta values. We set $\epsilon = 2$ for all the networks. The network in the upper left corner is a 'star' network. Here, the topological anonymity value is negative ($-5/6$) since there are only two degree sets and both contain breaches. Degree set D_1 has five members, but has no variance in the clustering coefficients. Degree set D_5 has only one member, so $|D_5| < \epsilon$. The network in the lower left corner is a fully connected network, where $ta = 0$. Here, there is not a node anonymity breach, but every node has an edge inference breach, i.e. the variance of the clustering coefficients is zero. Finally, the network on the right has a positive ta value. Intuitively, this results since every degree set has more than ϵ members and every degree set except one has variance in its edge connectivity structure.

The topological anonymity measure gives us a way to measure the level of anonymity of the connectivity structure of a network and provides users with insight into the obscuring power of the network topology.

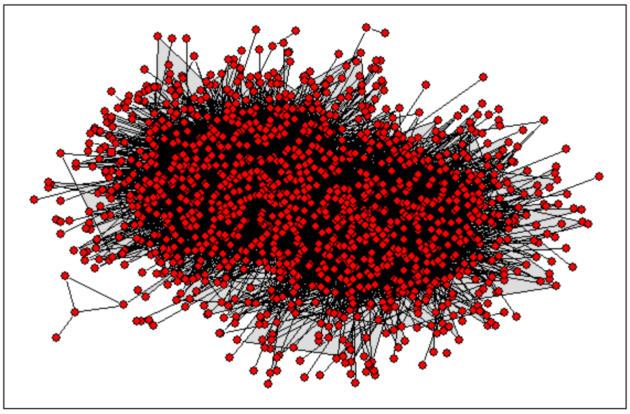

Figure 5. Political Blogs Network

5 Privacy evaluation of topological structures

In this section we analyze our measure in the context of two random graphs and one real world data set. Figure 4 shows an example of two random graphs. On the left is an Erdös-Renyl random graph where the degree of the nodes in the network follow a binomial distribution. On the right is an example of a scale free network in which the degree distribution follows a power law distribution. We chose a scale free network because the degree of nodes in many social networks has been shown to follow a power law distribution [2]. Both of these random networks have 100 nodes and an average degree of approximately 12. Our third data set, shown in Figure 5 is a political blog network data set [1]. Here links exist between blog sites based on a crawl of the front page of each blog. For this analysis, we only include blog sites with at least one link in the network. There are 1224 nodes and the average degree is 27.

Figure 6 shows the comparison of the topological anonymity measure for the synthetic networks and the political blog network. The x-axis shows ϵ values and the y-axis shows ta values for increasing values of ϵ. We see that for the scale free network topological anonymity decreases rapidly as $epsilon$ increases to only 4. Recall, that an $epsilon$ value of 4 means that 4 nodes must have the same degree for anonymity to exist. On the other hand, the Erdös-Renyl random graph maintains a high topological anonymity as ϵ increases. Given the underlying binomial distribution, this is not surprising. Finally, the political blog network has poor anonymity as well. Its degree distribution is closer to that of a scale free network. Looking at values of topological anonymity gives insight into how resilient our network is to privacy breaches. An important use of this measure is for evaluating how well different anonymity schemes work. For example, if the topological anonymity of a network is 0.1 and a proposed perturbation algorithm

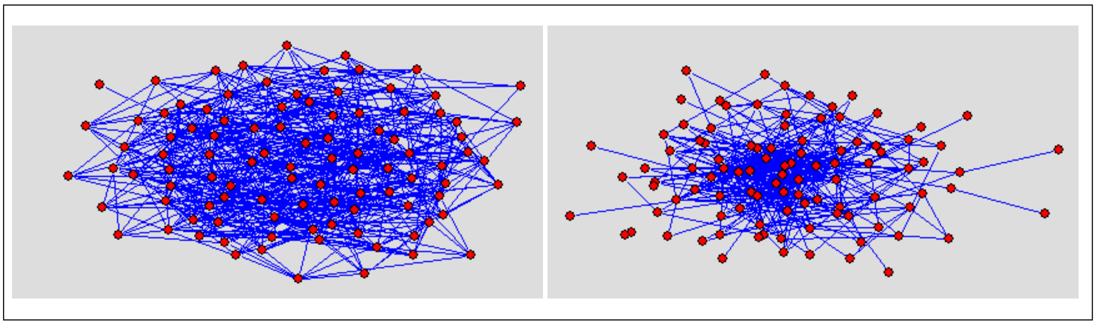

Figure 4. Random Networks - Left: Binomial distribution; Right: scale free distribution on right

increases the value to 0.2, users can evaluate whether the cost of perturbation is enough for the level of improvement in the ta value.

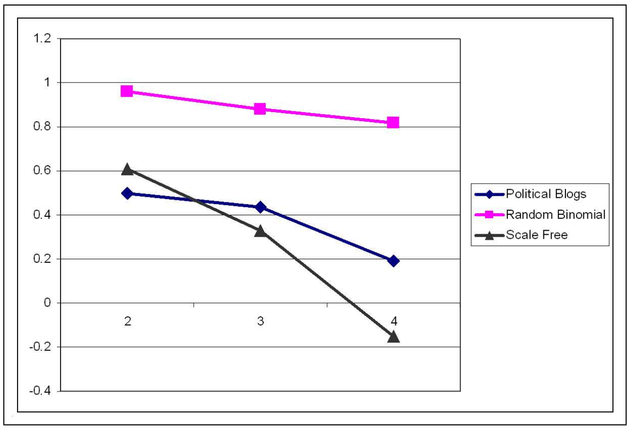

Figure 6. Comparison of Topological Anonymity

6 Conclusions and future directions

In this paper we present a metric (topological anonymity) that can be used to understand the level of anonymity in a social network based on the topology of the network. The metric is based on well known graph and social network topology measures, degree and clustering coefficient. Our notion of anonymity is based on two types of breaches, a node anonymity breach and an edge inference breach. If either of those local neighborhood breaches occur, the topological anonymity measure decreases.

We show that some topologies have more redundancy in them than others. For example, a random graph based on a binomial distribution of the degree of the node has a higher topological anonymity value than one based on a power law distribution. When analyzing graph structure, we see that a highly symmetric topology leads to edge inference breaches. A topology that is not symmetric enough leads to node identity breaches. Topological anonymity, ta attempts to balance this conflict by taking both node degree and clustering coefficient into account.

Now that we can identify the privacy issues, what algorithms should we use to efficiently make the network more resilient? What properties should be maintained in the network to get an approximation of the network that is meaningful for social network analysis applications and graph mining? How do we maintain the anonymity level when the topology of a given social network is changing? There are a lot of questions that still remain and a number of future directions that can be explored.

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 us election. In *The Proceedings of WWW-2005 Workshop on the Weblogging Ecosystem*, 2005.

[2] R. Albert, A. Barabsi, and H. Jeong. Scale-free characteristics of random networks: The topology of the world wide web, 2000.

[3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 181–190, New York, NY, USA, 2007. ACM Press.

[4] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. TCC*, 2006.

[5] F. Harary. *Graph theory*. Basic Books, 1994.

[6] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks, March 2007.

[7] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty*, 10(5):557–570, 2002.

[8] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. Cambridge University Press, Cambridge, 1994.

[9] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 6684, 1998.