Diversity, Homophily and the Risk of Node Re-identification in Labeled Social Graphs

Sameera Horawalavithana¹, Clayton Gandy¹, Juan Arroyo Flores², John Skvoretz² and Adriana Iamnitchi¹

¹ Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA sameeral@mail.usf.edu
² Department of Sociology, University of South Florida, Tampa, FL, USA

Abstract. Real network datasets provide significant benefits for understanding phenomena such as information diffusion or network evolution. Yet the privacy risks raised from sharing real graph datasets, even when stripped of user identity information, are significant. When nodes have associated attributes, the privacy risks increase. In this paper we quantitatively study the impact of binary node attributes on node privacy by employing machine-learning-based re-identification attacks and exploring the interplay between graph topology and attribute placement. Our experiments show that the population’s diversity on the binary attribute consistently degrades anonymity.

1 Introduction

Real graph datasets are fundamental to understanding a variety of phenomena, such as epidemics, adoption of behavior, crowd management and political uprisings. However, while the need for real social graphs and the supply of such datasets are well established, the flow of data from data owners to researchers is hampered by serious privacy risks: even when humans’ identities are removed, studies have proven repeatedly that de-anonymization is doable with high success rate [16]. Such de-anonymization techniques reconstruct user identities using third-party public data and the graph structure of the naively anonymized social network: specifically, the information about one’s social ties, even without the particularities of the individual nodes, is sufficient to re-identify individuals.

Many anonymization methods have been proposed to mitigate the privacy invasion of individuals from the public release of graph data [11]. Often the utility of an anonymized graph depends not only on preserving essential graph properties of the original graph, but also node attributes such as labels that identify node properties like gender or cheater status in online games [3].

However, the effects of node attributes on the risks of re-identifications are not yet well understood. While intuitively any extra piece of information can be a danger to privacy, a rigorous understanding of how topological and attribute properties affect the re-identification risks is needed. In cases such as information dissemination, node attributes may be correlated with local graph topology.
Our work assesses the additional vulnerability to re-identification attacks posed by node attributes in a graph. We consider exactly one binary attribute to understand the lower bound of the damage that node attributes inflict. We focus our empirical study on the interplay between topology and node attribute placement as a leverage point for re-identification. Because our focus is to understand in which conditions node re-identification is feasible, this study is independent of any anonymization technique. We apply machine learning techniques that use both topological and attribute information to re-identify nodes based on a common threat model. Our study involves real and synthetic graphs in which we control how node attributes are placed to mimic the characteristic homophily property of social networks [15].

Our empirical results show that the vulnerability to node re-identification depends on the population diversity with respect to the attribute considered. Using information about the distribution of attributes in a node’s neighborhood provides additional leverage for the re-identification process, even when attributes are binary. We quantify the relative importance of attribute-related and topological features in graphs of different characteristics.

2 Related Work

The availability of auxiliary data helps reveal the identities of anonymized individuals, as proven empirically in large privacy violation incidents [13, 6]. Similarly, in the case of graph de-anonymization attacks, information from an auxiliary graph is used to re-identify the nodes in an anonymized graph [16]. The quality of such an attack is determined by the rate of correct re-identification of the original nodes in the network. In general, de-anonymization attacks harness structural characteristics of nodes that uniquely distinguish them [11].

Efforts to incorporate node attribute information into de-anonymization attacks are few. Gong et al. [5] evaluate the combination of structural and attribute information on link prediction models. Attributes not present may be inferred through prior knowledge and network homophily. Qian et al. [17] apply link prediction and attribute inference to de-anonymization by quantifying the prior background information of an attacker using knowledge graphs. Ji et al. [12] showed a significant loss of anonymity when more node-attribute relations are shared between anonymized and auxiliary graph data. Specifically, they measure the entropy present in node-attribute mappings available for an attacker. As the entropy decreases, the graph loses node anonymity.

However, the success rate of a de-anonymization process is often reported in the literature as dependent on the chosen heuristic of the attack, which is typically designed with knowledge of the anonymization technique. Comparing the strengths of different anonymization techniques thus becomes challenging, if not impossible. Sharad [18] proposed a general threat model to measure the quality of a de-anonymization attack which is independent of any anonymization scheme. He proposed a machine learning framework that explores hidden invariants to re-identify nodes in the anonymized graphs.
The main aspects distinguishing this study from existing works are as follows:
i) In our work, we study the topological graph characteristics that provide resistance/vulnerability to node re-identification attacks based on machine learning techniques. Thus, structural perturbations for graph anonymization are irrelevant in our study. ii) To the best of our knowledge, this is the first work that quantifies the privacy impact of node attributes under an attribute attachment model biased towards homophily. iii) We analyze the interplay between graph structures and attribute information.

3 Methodology

Our main objective is to quantitatively estimate the vulnerability to reidentification attacks added by node attributes. Specifically, we ask: Given a graph topology, how much better does a node re-identification attack perform when the node attributes are included in the attack compared to when there is no node attribute information available to the attacker?

We are interested in measuring the vulnerability of a graph with node attributes in the absence of any particular anonymization technique on topology or node attributes. The intuition is that some graphs are inherently more private: for example, in a regular graph, nodes are structurally indistinguishable. Adding attributes to nodes, however, may contribute extra information that could make the re-identification attack more successful. For example, in a highly disassortative network (such as a sexual relationships network), knowing the attribute values (i.e., gender) of a few nodes will quickly lead to correctly inferring the attribute values of the majority of nodes, and thus possibly contributing to the re-identification of more nodes. Thus, the questions we address in this study also include: How does the distribution of node attributes affect the intrinsic vulnerability to a node re-identification attack?

To answer these question, we developed a machine learning-based reidentification attack inspired from that presented in [18]. We use the same threat model that aims at finding a bijective mapping between nodes in two different graphs. We mount a machine-learning based attack in which the algorithm learns the correct mapping between some pairs of nodes from the two graphs, and estimates the mapping of the rest of the dataset. As input data, we use both real and synthetic datasets (Section 4).

The threat model we consider is the classical threat model in this context: The attacker aims to match nodes from two networks whose edge sets are correlated. We assume each node is associated with a binary valued attribute, and this attribute is publicly available. For clarity, consider the following example: an attacker has access to two networks of individuals in an organization that represent the communication patterns (e.g., email) and friendship information available from an online social network. Individuals in the communication network are described by professional seniority (e.g., junior or senior), while individuals in the friendship network are described by gender. These graphs are structurally overlapping, in that some individuals are present in both graphs, even if their
identities have been removed. The attacker’s task is to find a bijective mapping between the two subsets of nodes in the two graphs that correspond to individuals present in both networks.

For training, we represent each node using a combination of two vectors that capture the node’s neighborhood degree distribution (NDD) and neighborhood attribute distribution (NAD). NDD is a vector of positive integers where $NDD^q_u[k]$ represents the number of $u$’s neighbors at distance $q$ with degrees between $50 \times k$ and $50 \times (k + 1) - 1$. We limit $q$ to 2. We use a bin size of 50, which was shown empirically [18] to capture the high degree variations of large social graphs. For each $q$, we use 21 bins, which would correspond to a largest node degree of 1050. All larger values are binned in the last bin. This binning strategy is designed to capture the aggregate structure of ego networks, which is expected to be robust against edge perturbation [18]. NAD is defined by $NAD^q_u[i]$ which represents the number of $u$’s neighbors at distance $q$ with an attribute value $i$. More details are provided in [10].

We generate examples for the training phase of the deanonymization attack by randomly picking node pairs from the sanitized ($G_{san}$) and the auxiliary ($G_{aux}$) graphs, respectively. In most cases, we have an unbalanced dataset with the degree of imbalance depending on the overlap parameter $\alpha$, where the majority is non-identical node pairs. We use reservoir sampling [7] to take $\ell = 1000$ balanced sub-samples from the population $S$, and the SMOTE algorithm [4] as an over-sampling technique for each sub-sample. Each sample is trained by a forest of 100 random decision trees that allows the algorithm to learn features. Gini-index is used as an impurity measure for the random forest classification. We measure the quality of the classifier on the task of differentiating two nodes as identical or not.

We use F1-score to evaluate the quality of the classifier. F1-score is the harmonic mean between precision and recall, typical metrics for prediction output of machine learning algorithms. For each data sample, we perform $5 \times 2$ cross-validations to evaluate the classifier and record the mean F1-score. We thus build two vectors of mean F1-scores, each of size $\ell = 1000$ (as described above), one for the attributed ($GS(LBL)$) and one for the unlabeled network topology ($GS$). An important aspect of these vectors is that they are related in the sense that the $i^{th}$ element in one vector represents the same sample as the $i^{th}$ element of the other vector. This is important for the pairwise comparison of the two mean F1-score vectors.

We perform a standard T-test on these two vectors and report the T-statistic value. The T-statistic value is a measure of how close to the hypothesis an estimated value is. In our case, the hypothesis is the prediction accuracy of the node identities in the unlabeled graph ($GS$) and the estimated value is the prediction accuracy in the labeled graph ($GS(LBL)$). Thus, a large T-statistic value implies a significantly better prediction accuracy of node identities in $GS(LBL)$ than in $GS$. In such cases, we can say that the network with node attributes is more vulnerable to node re-identification. This value serves as our statistical measurement to quantify the vulnerability cost of node attributes.
4 Datasets

Because our work is empirically driven, a larger set of test datasets promises a better understanding of the relations between vulnerability to re-identification attacks and the particular characteristics of the node attributes (such as fractions of attributes of a particular value or the assignment of attributes to topologically related nodes). In this respect, real datasets are always preferable to synthetic ones, as they potentially encapsulate patterns that are missing in the graph generative models.

However, relying only on real datasets has its limitations, due to the scarcity of relevant data (in this case, networks with binary node attributes) and the difficulty of covering the relevant space of graph metrics when relying only on available real datasets. Thus, in this work, we combine real networks with synthetic networks generated from real datasets.

4.1 Real Network Datasets

We used four networks with binary node attributes from publicly available datasets in three different contexts.

Polblogs [1] is an interaction network between political blogs during the lead up to the 2004 US presidential election. This dataset includes ground-truth labels identifying each blog as either conservative or liberal.

FB is a Facebook social network from University of Michigan in 2005 [20]. Node attributes such as dorm, gender, graduation year, and academic major are available. We chose two such attributes that could be represented as binary attributes: gender and occupation, whereby occupation identifies "student" or "faculty" individuals. From this dataset, we obtained thus two networks with the same topology but different node attribute distributions.

Amazon [14] is a bi-modal projection of categories in an Amazon product co-purchase network. Nodes are labeled as “book” or “music”, edges signify that the two items were purchased together.

As Table 1 shows, the networks generated from these datasets have different graph characteristics. For example, the density ($d$) of the graphs varies across two orders of magnitude, while degree assortativity oscillates between disassortative (for Amazon, $r = -0.06$ to assortative (as expected for social networks). All topologies except for Amazon have small average path length.

The metrics $p$ and $\tau$ shown in Table 1 are inspired from the synthetic node labeling algorithm used for generating synthetic graphs (and presented later), and they also show high variation across different networks. Intuitively, $p$ captures the diversity of attribute values in the node population (with $p = 0.5$ showing equal representation of the attributes) while $\tau$ captures the homophily phenomenon (that functions as an attraction force between nodes with identical attribute values). The homophilic attraction metric $\tau$ varies between 0.37 in FB (thus, no higher than chance preference) to 0.99 in Amazon (books are purchased together with other books). The diversity metric $p$ varies from 8% female representation in the FB dataset to a near perfect blog category representation in the
Polblogs (where \( p = 0.48 \)). We only consider \( p \) as the minimum proportion of two node groups due to the symmetric nature of attributes in our experiments. This wide variation in graph metrics values is what motivated our choice for these network datasets.

Table 1: Graph properties of the real network datasets. All graphs are undirected, and nodes are annotated with a binary valued attribute (R or B). \( p \) and \( \tau \) are the measured parameter values of the attraction model, which quantitatively describe the diversity and homophily in the network. \( \bar{d} \) represents graph density, \( C \) is transitivity, \( (r) \) is degree-assortativity, and \( \kappa \) is the average path length.

| Network   | \( |N| \) | \( |E| \) | \( p \) | \( \tau \) | \( \bar{d} \) | \( C \) | \( r \) | \( \kappa \) |
|-----------|------|------|------|------|------|------|------|------|
| Polblogs  | 1224 | 16718| 0.02 | 0.22 | 0.22 | 2.49 |
| (party)   | 48   | 48   | 0.48 | 0.34 | 0.02 | 0.22 | 0.22 | 2.49 |
| FB        | 30147| 1176516| 0.0026 | 0.13 | 0.115 | 3.05 |
| (gender)  | 92.2 | 7.8  | 90.5 | 0.2  | 9.3  | 0.08 | 0.37 | 0.13 |
| (occupation) | 17.5 | 22.5 | 72   | 9    | 19.6 | 0.22 | 0.46 | 0.13 |
| Amazon    | 305561 | 853326 | 0.18 | 0.99 | 0.18 \( \times 10^{-5} \) | 0.21 | 0.06 | 17.42 |

4.2 Synthetic Graphs

In order to be able to control graph characteristics and node attribute distributions, we also generated a number of synthetic graphs comparable with the real datasets just described. The graph generation included two aspects: topology generation, for which we opted for ERGMs, and node attribute assignments, for which we implemented the technique proposed in [19].

Varying Topology via ERGMs Exponential-family random graph models (ERGMs) or p-star models [9] are used in social network analysis for stipulating, within a set structural parameters, distribution probabilities for networks. Its primary use is to describe structural and local forces that shape the general topology of a network. This is achieved by using a selected set of parameters that encompass different structural forces (e.g., homophily, degree correlation/assortativity, clustering, and average path length). We used \( R \) and the statnet suite [8] which contains several packages for network analysis to generate graphs with similar structural characteristics as the real-world network datasets presented above. We focused on three structural graph metrics: clustering coefficient, average path length, and degree correlation/assortativity.

Synthetic Labeling A simple model that parameterizes a labeled graph with a tendency towards homophily (ties disproportionately between those of similar
attribute background) is an “attraction” model [19]. In the basic case of a binary attribute variable and a constant tendency to inbreed, two parameters, \( p \) and \( \tau \), both in the (0,1) interval, characterize the distribution of ties within and between the two groups. The first is the proportion of the population that takes on one value of the attribute (with \( 1 - p \), the proportion taking on the other value). The second parameter, the inbreeding coefficient or probability, expresses the degree to which a tie whose source is in one group is "attracted" to a target in that group. When \( \tau = 0 \), there is no special attraction and ties within and between groups occur in chance proportions. When \( \tau > 0 \), ties occur disproportionately within groups, increasing as \( \tau \) approaches 1. Given a total number of ties, values for \( p \) and \( \tau \) determine the number of ties/edges that are between groups, namely, \( \delta = |E| \times 2 \times (1 - \tau) \times p \times (1 - p) \).

In the process of generating synthetic node attributes, we first randomly assign two arbitrary values (i.e., R and B) as labels to all the nodes in the graph for a given \( p, 1 - p \) split. Then, we draw an R node and a B node at random and swap labels if it would decrease the number of R–B ties. This process would converge when the total number of cross-group ties is reduced to \( \delta \) for a particular value of \( \tau \). Table 2 presents the graph characteristics of the synthetically generated labeled graphs.

| Network | ERGM | \( d \) | \( C \) | \( r \) | \( \kappa \) | \( |S| \) (millions) |
|---------|------|------|------|------|------|----------------|
| polblogs | dc | 0.02 | 0.03 | 0.08 | 2.52 | 5.5 |
|          | cc | 0.02 | 0.33 | 0.02 | 2.69 | 13.1 |
|          | apl | 0.02 | 0.10 | -0.06 | 2.49 | 11.5 |
| FB     | dc | 0.003 | 0.02 | 0.12 | 3.28 | 38.4 |
|          | cc | 0.002 | 0.20 | 0.12 | 3.52 | 39.9 |
|          | apl | 0.002 | 0.20 | 0.12 | 3.64 | 38.9 |
| Amazon | dc | 1.82E-5 | 0.37 | -0.06 | 11.86 | 43.7 |
|          | cc | 1.82E-5 | 0.40 | -0.06 | 13.52 | 72.5 |
|          | apl | 1.82E-5 | 0.39 | -0.06 | 13.47 | 74.3 |

5 Empirical Results

Our objective is not to measure the success of re-identification attacks on original datasets in which node identities have been removed: it has been demonstrated long ago [2] that naive anonymization of graph datasets does not provide privacy. Instead, our objective is to quantify the exposure provided by node attributes on
top of the intrinsic vulnerability of the particular graph topology under attack. Our first guiding question is thus: *How much risk of node re-identification is added to a network dataset by its binary node attributes?*

### 5.1 The Vulnerability Cost of Node Attributes

Table 3 reports the quality of node re-identification in the original graph topology GS with the quality of node re-identification in the same topology augmented with node attributes GS(LBL). As expected, re-identification is (generally) better when node attributes are used, yet that difference is relatively small, between 0.01 and 0.07.

![Table 3: The quality of node re-identification is presented using 5×2 cross-validation F1-scores. Mean and std. deviation of F1-scores are calculated over \(\ell = 1000\) re-identification attacks as shown for real network datasets on GS and GS(LBL) along with the T-statistic.](image)

<table>
<thead>
<tr>
<th>Network</th>
<th>GS F1-score</th>
<th>GS (LBL) F1-score</th>
<th>T-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>stdv</td>
<td>mean</td>
</tr>
<tr>
<td>Polblogs (party)</td>
<td>0.8127</td>
<td>0.0174</td>
<td>0.8813</td>
</tr>
<tr>
<td>FB (gender)</td>
<td>0.8434</td>
<td>0.0171</td>
<td>0.8502</td>
</tr>
<tr>
<td>FB (occupation)</td>
<td>0.8664</td>
<td>0.0148</td>
<td>0.8723</td>
</tr>
<tr>
<td>Amazon (category)</td>
<td>0.5996</td>
<td>0.0344</td>
<td>0.6333</td>
</tr>
</tbody>
</table>

More interestingly, the success of re-identification attack is different for the same topology when the binary node attributes are placed in the network following different patterns. As an example, the *occupation* attribute significantly improves the attacker’s performance in the *FB* network compared to the *gender* attribute: the T-statistics value increases from 16.11 in gender *FB* to 58.48 in occupation *FB*. We thus formulate a new question: *What placement of attributes onto nodes reveal more information?*

### 5.2 Diversity Matters, Homophily Not

To understand how the placement of attribute values on nodes affects vulnerability, we generate synthetic node attributes in a controlled manner. By varying \(p\) (the diversity ratio) and \(\tau\) (the bias of nodes with same-value attributes to be connected by an edge), we can study the effect of these parameters on node re-identification.

Figure 1 presents the T-statistics of the F1-scores for node re-identification attacks on the original topology with and without node attributes. In addition, Figure 1 also presents results on various synthetically generated and labeled networks as described in Section 4.2.

We observe three phenomena: First, \(p\) is positively correlated with the T-statistic value that measures the re-identification impact of attributes. That is,
the higher diversity (i.e., the larger $p$), the more vulnerable to re-identification the labeled nodes become on average. In a highly skewed attribute population, while the minority nodes will be identified quicker due to node attributes, the majority remains protected. On the other hand, when $p = 0.5$, a network has two equal-sized sets of nodes where each set takes one of two attribute values. This is explained by the fact that the NAD feature vector captures more diverse information in the attributes of neighbors when $p$ is larger. This is also the explanation for why the node attributes contribute more to vulnerability in the FB dataset under the attribute *occupation*, which has a relatively high diversity ($p = 0.22$) than *gender* ($p = 0.08$). Note that the effect of $p$ on the added vulnerability remains consistent across all topologies tested.

The second observation is that there is no visible pattern on how $\tau$ influences the vulnerability added by binary node attributes. While this is disappointing from the perspective of story telling, it is potentially encouraging for data sharing, as it suggests that datasets that record homophily (or influence, the debate is irrelevant in this context) do not have to be anonymized by damaging this
pattern. As a specific example, the privacy of a dataset that records an information dissemination phenomenon could be provided without perturbing the cascading-related ties.

The third observation is on the relative effect of the topological characteristics on the added vulnerability. Amazon network is an order of magnitude sparser than the other datasets considered, and thus the topological information available to the machine learning algorithm is limited. In this situation, the addition of the attribute information turns out to be very significant: the T-statistic values for these datasets are significantly larger than for the other datasets, with values over 100 in majority of cases.

Another topological effect is noticed when comparing the real Amazon and FB topologies with the ERGM-generated ones in Figure 1b and Figure 1c: the node attribute contributes more to the original topology in FB than it does to the synthetic topologies, while in Amazon the opposite effect is observed. This may be due to the relative contribution of the original topological particularities, and rigorously understanding these correlations is a topic for future work.

6 Summary and Discussions

This work quantifies the increase in vulnerability to node re-identification attacks due to binary node attributes. We study how node vulnerability is affected by network properties and the placement of node attributes. We find that a population’s diversity on the binary attribute consistently degrades anonymity and increases vulnerability. Diversity means a more even distribution of the binary attribute which produces a more varied set of neighborhood distributions that a particular node may exhibit. Consequently, nodes are more easily distinguished from one another by virtue of their differing neighborhood distributions of labels.

One puzzle remains. There is no consistent discernible impact of homophily, as measured by the inbreeding coefficient, on vulnerability. Our procedure for investigating the impact of homophily simply involves swapping labels without disturbing ties. Therefore, both local and global (unlabeled) topologies remain constant as we decrease the number of cross-group ties to achieve a target value implied by a particular inbreeding coefficient for a given proportional split along the binary attribute. This procedure disturbs the local labeled topology but because the machine learning attack uses information from that local topology it apparently can adapt to the changes and make equally successful predictions regardless of the value of the inbreeding coefficient. Perhaps that is why many different factors in attacks on the labeled graphs have some degree of responsibility for success and, no relatively small subset gets the lion’s share of the credit.

Acknowledgement

This research was supported by the U.S. National Science Foundation under Grant IIS 1546453.
References
