

How friendship links and group memberships affect the privacy of individuals in social networks

Elena Zheleva, Lise Getoor
Department of Computer Science
College Park, Maryland 20742, USA
{elena,getoor}@cs.umd.edu

July 1, 2008

Abstract

In order to address privacy concerns, many social media websites allow users to hide their personal profiles from the public. In this work, we show how an adversary can exploit a social network with a mixture of public and private user profiles to predict the private attributes of users. We map this problem to a relational classification problem and we propose a simple yet powerful model that uses group features and group memberships of users to perform multi-value classification. We compare its efficacy against several other classification approaches. Our results show that even in the case when there is an option for making profile attributes private, if links and group affiliations are known, users' privacy in social networks may be compromised. On a dataset from a well-known social-media website, we could easily recover the sensitive attributes for half of the private-profile users with a high accuracy when as much as half of the profiles are private. To the best of our knowledge, this is the first work that uses link-based and group-based classification to study privacy implications in social networks. We conclude with a discussion of our findings and the broader applicability of our proposed model.

1 Introduction

A number of social media and social network websites, such as Facebook, Orkut and Flickr, allow their participants to set the privacy level of their online profiles and to disclose either some or none of the attributes in their profiles. Not surprisingly, some users are more open to sharing personal information than others and they disclose more attributes on their profiles. For example, some people feel comfortable displaying personal attributes such as age, political affiliation or location, while others do not. Most social-media users utilize the social network underlying the service by forming friendship links and affiliating with groups of interest. While a person's profile may remain private, the friendship

links and group affiliations are often visible to the public. Unfortunately, these friendships and affiliations leak information; in fact, as we will show, they can leak a surprisingly large amount of information.

Being able to control who learns their personal information is an important aspect for users who agree to participate in online social networks. For that reason, most social media websites allow users to hide their profiles from the general public. However, we think it is also important to be able to hide friendship lists and group memberships to ensure that they do not imply something that users did not intend to disclose. For example, in Facebook many users choose to set their profiles to private, so that noone but their friends can see their profile details. Yet, fewer people hide their friendship lists and even if they do, their friendship links can be found through the backlinks from their public-profile friends. Similarly with groups – even if a user hides his profile, his participation in a public group is shown on the group’s membership list. In order to ensure the privacy level that users desires, it is important that the users are aware of the privacy breaches that their friendship links and group participation entails. It is also important that the social media website provider protects its users against undesired eavesdropping by informing them of the possible privacy breaches and providing them the means to be in full control of their private data.

The problem we consider is *sensitive attribute inference* in social networks: inferring the private information of users given a social network in which some profiles are public and all links and group memberships are exposed. To the best of our knowledge, our work is the first one to look at this problem, to map it to a classification problem in network data with groups. More concretely, the data that we consider consists of entities that have links among each other and participate in groups together. The classes of some entities are known (public profiles), and the classes of the rest need to be inferred (private profiles). The entities are assumed to have no other attributes except the class value.

Here, we propose five classification models for inferring the sensitive attribute. The first simple baseline method ignores link and group information. Two of the models are link-based classification models; one considers aggregates over the friends of the user and the second one builds upon an idea borrowed from stochastic blockmodeling, that user interactions can be explained by the clusters to which they belong. The fourth model is group-based classification, an approach which takes into consideration the class values of entities that participate in the same groups as the object we are trying to classify. Our group-based classification model contains two main parts. First, it selects the groups that are relevant to the classification task. Next, it uses the groups as entity features: it classifies the entities by training a classifier based on the observed, public-profile information, and using it to predict unobserved, private attribute values. The fifth model is a simpler version of the group-based classification in which all the available groups in the classification process.

We evaluate the models on multi-value and binary classification tasks using sample datasets from three well-known social media websites - Flickr, Dogster

and BibSonomy¹. Our results show that the group-based classification achieves significantly better accuracy than the models that ignore the group information.

Our contributions include the following:

- We identify a number of novel privacy attacks in social networks with a mixture of public and private profiles
- We propose a general framework for group-based node classification in social networks with a large number of affiliation groups
- We show that automatic reduction of the large number of groups and using only the ones that are relevant to the classification task at hand is beneficial
- We show the privacy implications of publicly affiliating with groups in social networks and discuss how our study affects anonymization of social networks
- We evaluate our group-based classification approach on hard classification tasks in three social media datasets

Our results show that even in the case when there is an option for making profile attributes private, if links and group affiliations are known, users' privacy in social networks is illusionary at best.

2 Preliminaries

Predictive modeling in network data usually relies either on supervised or unsupervised learning. In the last decade, there has been a growing interest in supervised classification that relies not only on the object attributes but also on the attributes of the objects it is linked to, some of which may be unobserved [10, 14, 13]. Link-based classification (also known as relational or collective classification) in network data, such as social networks, relies on autocorrelation, the property that makes the classes of linked objects correlated with each other. For example, political affiliations of friends tend to be similar, a person communicating with criminals may be a criminal, etc.

The goal of unsupervised learning or clustering is to group objects together based on their similarity. In social networks, clusters can be found based on attribute and/or structural information. For example, Neville and Jensen [13] describe how autocorrelation in relational data is sometimes caused by the presence of such hidden clusters or groups in the data which influence the attributes of the group members. They use a spectral clustering method based on node links in the data to discover groups, and then use the groups to classify the nodes. Unlike other clustering methods for relational data, their method assumes that groups do not overlap. Airoidi et al. [2] study mixed-membership

¹At <http://www.flickr.com>, <http://www.dogster.com> and <http://www.bibsonomy.org/>

clustering of relational data to predict protein function. It is assumed that the cluster assignment is related to the node attribute value in question.

In contrast to these approaches, we are interested in classifying nodes when group membership is explicitly given and only a subset of the groups is related to the node attribute in question. This is different from the case where groups need to be detected because explicit groups can represent a latent common interest that neither attribute nor structural information contains. We propose a node classification method that makes use of explicit groupings with member-set overlaps, and it distinguishes groups that are relevant to classification based on group features such as name, size, link density, homogeneity, etc.

2.1 Motivation

By participating in a social network, people are vulnerable to disclosing personal data they may have not intended to disclose because some network members are more open to sharing their personal information than others. Our work shows that even if users are very conservative in displaying personal information, the relationships they form and the groups they join can help an adversary infer personal information that they may not have intended to disclose.

According to Li et. al. [6], there are two types of privacy attacks in data: *identity disclosure* and *attribute disclosure*, and identity disclosure often leads to attribute disclosure. Identity disclosure occurs when the adversary is able to determine the mapping from a record to a specific real-world entity (e.g. an individual). Attribute disclosure occurs when an adversary is able to determine the value of a user attribute that the user intended to stay private.

Until recently, the literature on privacy preservation considered only single-table data, in which the rows represent i.i.d. records, and the columns represent record attributes [1, 4, 6, 9, 12, 15]. Real-world data is often relational, and records may be related to one another or to records from other tables. Relational data poses new challenges to preserving the privacy of individuals [3, 5, 8, 11, 18]. For example, in graph data, there is a third type of disclosure attack: *link re-identification* [18]. Link re-identification is the problem of inferring that two entities participate in a particular type of sensitive relationship or communication. If one anonymizes the data naïvely by removing personal attributes and replacing them with a random identifier, it still is possible to identify individuals based on their subgraph structure [3, 5, 8]. It is also possible to link records in anonymized data to external relational data sources to disclose attribute values [11]. Our work is complementary in that we assume that the identities of people are known but the value of the sensitive attribute of some of them is not directly available. We propose several simple models for inferring the hidden sensitive attributes using the observed attributes, link and group information in a single data source.

All the privacy attacks mentioned above are meant to show that more sophisticated anonymization techniques are necessary. It is important to be aware of the different possible privacy attacks in order to guide these anonymization techniques. The challenge of anonymizing graph data lies in understanding the

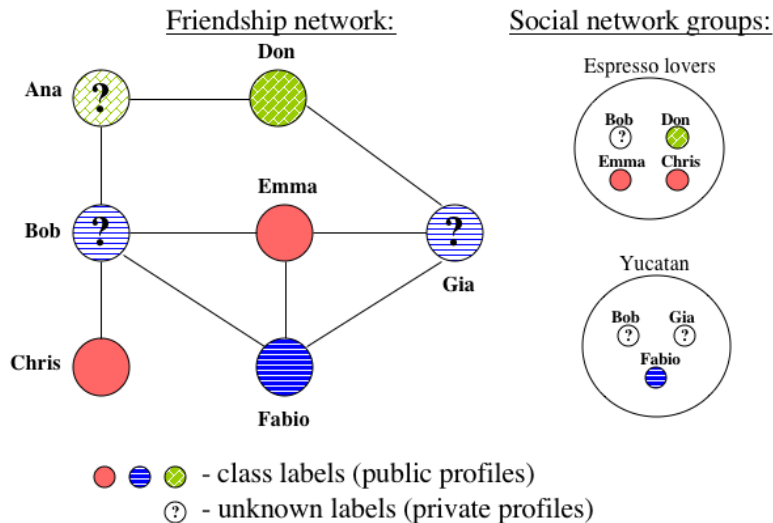


Figure 1: Toy instance of the data model.

rich dependencies in the data and removing sensitive information which can be inferred by direct or indirect means. Here, we show an attribute-disclosure attack in data which is meant to be partially private. We look at the attribute disclosure problem as a classification problem and we use friendship links, group affiliation and public attribute values as its features.

Based on the group-based privacy attack, we identify a fourth type of privacy breach in relational data, *group membership disclosure*: whether a person affiliates with a group relevant to a sensitive-attribute classification. We conjecture that hiding group memberships is important in preserving the privacy of individuals and their personal data. A group membership disclosure can lead to an attribute disclosure.

3 Data model

We represent the relational data as a graph $G = (V, E, H)$, where V is a set of n nodes of the same type, E is a set of directed edges, and H is a set of groups that a node can belong to. $e_{i,j} \in E$ represents a relationship between the nodes v_i and v_j . We describe a group as a hyper-edge $h \in H$ among all the nodes who belong to that group; $h.U$ denotes the set of users who are connected through hyper-edge h and $v.H$ denotes the groups that node v belongs to. Similarly, $v.F$ is the set of nodes that v has connected to: $v_i.F = \{v_j | \exists e_{i,j} \in E\}$. A group can have a set of properties $h.T$.

We assume that each node v has a sensitive attribute which is either observed or hidden in the data. A *sensitive attribute* is a personal attribute, such as age, political affiliation or location, which some users in the social network are willing

to disclose publicly while others keep private. A sensitive attribute value can take on one of a set of possible values $\{a_1 \dots a_m\}$. A *user profile* has a unique id with which the user forms online relationships and participates in groups. Each profile is associated with a sensitive attribute, either observed or hidden. Based on that, there are two types of profiles: private and public. A *private profile* is one for which the sensitive attribute value is unknown, and a *public profile* is the opposite: a profile with an observed sensitive attribute value. We refer to the set of nodes with private profiles as the *sensitive set* of nodes V_s , and to the rest as the *observed set* V_o . The adversary’s goal is to predict V_s , the sensitive attributes of the private profiles.

A group of users has a distribution over the observed sensitive attribute values, and this is the information one can use to predict the sensitive attribute values of the users whose values are unobserved. We are interested in the case where nodes have no other attributes except the sensitive attribute because when a user profile is private, no attributes are available for that user. This case is different from a traditional classification case where a classifier depends on the other attributes of the node. Without attributes, it is not possible to train a standard classifier such as Naïve Bayes or logistic regression to predict the sensitive attribute using other attributes of the same node. Therefore, the algorithm has to rely on links, groups and domain knowledge.

As a running example network, we consider the social network presented in Figure 1 which contains the friendship network and the groups of interest. Chris, Don, Emma and Fabio are displaying their attribute values publicly, while Ana, Bob and Gia are keeping theirs private. Emma and Chris have the same sensitive attribute value. Users are linked by a friendship link, and in this example they are reciprocal. There are two groups that users can participate in: the "Espresso lovers" group and the "Yucatan" group. While affiliating with some groups may be related to the sensitive attribute, affiliating with others is not. For example, if the sensitive attribute is a person’s country of origin, the "Yucatan" group may be relevant. Ideally, we would like to be able to reason about these probabilistically.

4 Private-attribute inference models

Online communities allow very diverse people to connect to each other and form relationships that transcend gender, religion, origin and other boundaries. As this happens, it seems harder to utilize the complex interactions in online social networks for predicting user attributes because the correlation of friends’ attributes does not necessarily hold. Attribute disclosure occurs when the adversary is able to infer the sensitive attribute of a real-world entity accurately.

An individual’s attribute value could be looked at as a representative of the social network attribute distribution, as well as a representative of its friendship network attribute distribution, or a representative of each group that he or she belongs to. The basic intuition behind the models we present is that the sensitive-attribute distributions of the group and friendship circles of a person

are more informative about his/her sensitive attribute than the overall sensitive-attribute distribution.

The problem of sensitive attribute prediction is to infer the hidden sensitive values $V_s.A$ conditioned on the observed sensitive attribute values, links and group membership in graph G . We assume that the adversary can apply a probabilistic model F for predicting the hidden sensitive attribute values, and he can combine the given graph information in various ways as we discuss next. The node prediction of each model is:

$$v_s.\hat{a}_F = \operatorname{argmax}_{a_i} P_F(v_s.a_i; G).$$

where $P_F(v_s.a_i; G)$ is the probability that the sensitive attribute value of node v is a_i according to model F and the observed part of graph G .

4.1 Privacy attack in the absence of links and groups

In the absence of relationships and groups, we assume that the given information for an individual is the overall dataset distribution of the sensitive value. This approach uses the distribution of values in the nodes. More formally, according to the model based on node information F_V , the probability of a sensitive attribute value can be estimated as the fraction of observed users who have that sensitive attribute value:

$$P_{F_V}(v_s.a_i; G) = P(v_s.a_i|V_o.A) = \frac{|V_o.a_i|}{|V_o|}.$$

The adversary using model F_{V_o} picks the most probable attribute value which in this case is the overall mode of the multinomial attribute distribution. In our toy example, the adversary would predict that Ana, Bob and Gia have the same attribute value as Chris and Emma. One problem with this approach is that if there is a sensitive attribute value that is predominant in the observed data, it will be predicted for all users with private profiles. Next, we look at using friendship information for inferring the attribute value.

4.2 Privacy attacks using links

Link-based models take advantage of *autocorrelation*, the property that the attribute values of linked objects are correlated. One example of autocorrelation is people who are friends sharing some of the same personal features. This relates to the proverb found in many cultures "Tell me who your friends are, and I'll tell you who you are." Our link-based setup considers network data which consists of linked nodes which have no other attributes besides the class label. Figure 2 (b) shows a graphical representation of the link-based classification model.

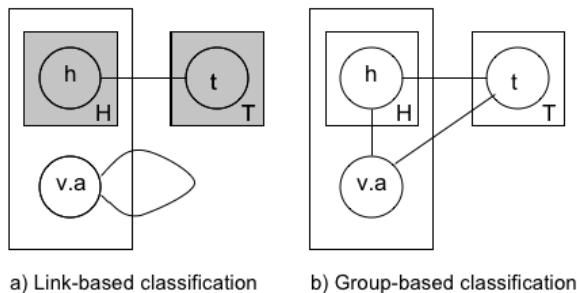


Figure 2: Graphical representation of the models. Grayed areas correspond to variables that are ignored in the model.

4.2.1 Friend-aggregate model

The nodes and their links give a graph structure in which one can identify circles of close friends. For example, the circle of Bob’s friends, is the set of users that he has links to: $Bob.F = \{Ana, Chris, Emma, Fabio\}$. The friend-aggregate model makes use of *autocorrelation*, and looks at the sensitive attribute distribution amongst the friends of the person under question. According to the model based on node and link information $F_{V,E}$, the probability of the sensitive attribute value can be estimated by:

$$P_{F_{V,E}}(v_s.a_i; G) = P(v_s.a_i | V_o.A, E) = \frac{|V'_o.a|}{|V'_o|}$$

where $V'_o = \{v_o \in V_o | \exists (v_o, v_s) \in E\}$.

Again, the adversary using model $F_{V,E}$ picks the most probable attribute value (i.e., the mode of the friends’ attribute distribution). This model is similar to the weighted-vote relational-neighbor procedure with unweighed edges described in [10]. In our toy example (Figure 1), Bob would pick the same value as Emma and Chris, Ana the same label as Don, and Gia will be undecided between Don’s, Emma’s and Fabio’s label. One problem with this method is that if the class distribution is skewed, then highly-represented class labels may take over the predicted labels. Another problem is the one in which a person’s friends are very diverse, as is Gia’s case.

4.2.2 Blockmodeling-based model

The basic idea behind *stochastic blockmodeling* is that users form natural clusters or blocks, and their interactions can be explained by the blocks they belong to [17, 2]. In particular, the link probability between two users is the same as the link probability between their corresponding blocks. If sensitive attribute values separate users in blocks, then based on the observed interactions of a private-profile user with public-profile users, one can predict the most likely block that the user belongs to and thus discover the attribute value. Let block

B_{a_i} denote the set of public-profile users who have attribute value a , and $\lambda_{i,j}$ the probability that a link exists between users in block i and users in block j . Thus, λ_i is the vector of all link probabilities between block i and each block B_1, \dots, B_m . Similarly, let the probability of a link between a single user v and a block j be $\lambda(v)_j$ with $\lambda(v)$ being the vector of link probabilities between v and each block. To find the probability that a private-profile user belongs to a particular block, the model looks at the maximum similarity between the interaction patterns (link probability to each block) of the node in question and the overall interactions between blocks. After finding the most likely block, the sensitive attribute value is predicted. The probability of an attribute value using the blockmodeling-based model F_B is estimated by :

$$P_{F_B}(v_s.a_i; G) = P(v_s.a_i | V_o.A, E, \lambda) = \frac{1}{Z} \text{sim}(\lambda_i, \lambda(v))$$

where $\text{sim}()$ could be any vector similarity function and Z is a normalization factor. We compute maximum similarity using the minimum L2 value. This model is similar to the class-distribution relational classifier described in [10] when the weight of each directed edge is inversely proportionate to the size of the class of the receiving node.

4.3 Privacy attack using group memberships

Social networks offer a very rich structure through the group memberships of users. Groups offer a broad perspective on a person, and it may be possible to use the group affiliations in order to achieve better classification. All individuals in a group are bound together by some observed or hidden interest that they share. One can think of groupmates as users to whom one is implicitly linked to and again apply the idea of autocorrelation.

If a user belongs to only one group (as it is Gia’s case in the toy example), then it is straightforward to use the aggregate, e.g., the mode, of her groupmates’ labels, similar to the friend-aggregate model to assign a label to her. This problem becomes more complex when there are multiple groups that a user belongs to, and their distributions suggest different values for the sensitive attribute. Moreover, some of the groups may be irrelevant to the classification task at hand and ideally, we would like to discard them. For example, the group ”Yucatan” may be relevant for finding where a person is from but ”Espresso lovers” may not be.

A group is either relevant to the classification task at hand or not. To select the relevant groups, one can apply standard feature selection criteria [7]. If there are N groups, the number of candidate group subsets is 2^N , and finding an optimal feature subset is intractable. Similar to pruning words in document classification, one can prune groups based on their properties and evaluate their predictive accuracy. Example group properties include density, size, homogeneity. Smaller groups may be more predictive than large groups, and groups with high homogeneity may be more predictive of the class value.

For example, if the classification task is to predict the country that people are from, cultural groups may be more relevant to this task. A group which is more homogeneous in relation to the class labels of its members is more likely to be relevant. For example, a group in which 90% of the people are from the same country is more likely to be predictive of the country class label. One way to measure group homogeneity is by computing the entropy of the group. A group h can be looked at as a discrete random variable that can take on the possible node-class values of its observed node members.

$$Entropy(h) = - \sum_{i=1}^m p(a_i) \log_2 p(a_i)$$

where m is the number of possible node class values and $p(a_i)$ is the fraction of observed members that have class value a_i :

$$p(a_i) = \frac{|h.V.a_i|}{|h.V|}.$$

For example, the group "Yucatan" has an entropy of 0 because only one attribute value is represented there, therefore its homogeneity is very high. Another group property is the percent of public profiles in it. This shows how confident we are in the computed group entropy which is estimated based on the public profiles.

In the most basic case, group features are ignored and all groups are considered to be relevant to the classification task. This case can be used as a straw man to check whether selecting groups based on their relevance is worthwhile.

Algorithm 1 Group-based classification model

```

1: Set of relevant groups  $H_{relevant} = \emptyset$ 
2: for each group  $h \in H$  do
3:   if  $isRelevant(h)$  then
4:      $H_{relevant} = H_{relevant} \cup \{h\}$ 
5:   end if
6: end for
7:  $trainClassifier(f, V_o, H_{relevant})$ 
8: for each sensitive node  $v \in V_s$  do
9:    $v.\hat{a} = f(v.H_{relevant})$ 
10: end for

```

The group-based classification approach contains three main steps as Algorithm 1 shows. In the first step, the algorithm performs feature selection: it selects the groups that are relevant to the node classification task. This can either be done automatically or by a domain expert. Ideally, when the number of groups is high, the feature selection should be automated. For example, the function $isRelevant(h)$ can return *true* if the entropy of group h is low. In the second step, the algorithm learns a global function f , e.g., trains a classifier, that takes the relevant groups of a node as features and returns the node class

value. This step uses only the nodes from the observed set because their sensitive attributes are known. Each node v is represented as a binary vector where each dimension corresponds to a unique group: $\{groupId : isMember\}$, $v.a$.

Only memberships to relevant groups that v belongs to are considered and $v.a$ is the class coming from a multinomial distribution which denotes the sensitive-attribute value. In the third step, the classifier returns the predicted sensitive attribute for each private profile. Figure 2 (b) shows a graphical representation of the group-based classification model.

5 Experiments

We evaluated each of the proposed models and assessed their accuracies. For the group models, we used an implementation of SVM with multi-value classification [16].

5.1 Data description

For our data experiments, we look at three diverse online communities, namely the photo-sharing website Flickr, the dog online social network Dogster and the social bookmarking system BibSonomy. For Flickr, we concentrated on the problem of predicting the country of a person from 55 possible values. For Dogster, we predicted the breed group of each dog from 7 possible values. We also experimented with a publicly available dataset from the social bookmarking system Bibsonomy, and we performed binary classification to detect malicious behaviour, i.e., whether a user is a spammer or not.

Flickr is a photo-sharing community in which users can display their photographs, comment on other users' photos, create directed friendship links, form and participate in groups of common interest. Users have the choice of providing personal information on their profiles, such as gender, marital status and location. We collected a snowball sample of 14,451 users from it. To resolve the location attributes (which users enter manually, as opposed to choosing them from a list), we used a two-step process. In the first step, we used Google Maps API² to find and unify the latitude and longitude of each user location. In the second step, we mapped the latitude and longitude back to a country location using the reverse-geocoding capabilities of GeoNames³. At the end, 34% of the users had no resolved country location. We considered only the users from countries which had at least 10 representatives. The sample contained people from 55 countries. There were 9179 users, 47754 groups with at least 2 members, 941677 directed links and 1486689 group memberships. The largest group has 4527 users.

Dogster is a pet social networking website where dog owners can create profiles describing their dogs, and to post and share information that includes photos and personal characteristics, as well as membership in community groups.

²At <http://code.google.com/apis/>.

³At <http://www.geonames.org/export/>.

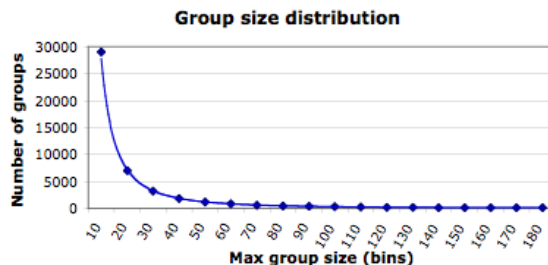


Figure 3: The group size in our Flickr sample follows a power-law distribution. The long tail is not shown on the figure. The maximum group size is 4527.

Members also maintain links to dog friends and family members. Our second dataset contains a random sample of 10,000 instances from Dogster. The dogs that do not participate in any groups were removed from the sample. The remaining 2,632 dogs participate in 1042 groups with at least two members each. Each dog has a breed such as *golden retriever* or *beagle*. Each breed belongs to a broader type set. In our dataset, there were mostly *toy* dogs (749). The other major breed categories were *working* (268), *herding* (202), *terrier* (232), *sporting* (308), *non-sporting* (225), *hound* (152) and *mixed dogs* (506).

The third dataset is publicly available from the ECML PKDD 2008 Discovery Challenge website⁴. It contains data from the social bookmarking website BibSonomy, in which users can tag bookmarks and publications. Even though BibSonomy allows users to join groups of interest, the dataset did not contain this information. Therefore, we consider each tag placed by a person to be a group that a user belongs to. We considered tag instances for both bookmarks and publications, and converted them all to lower case. There are no links between users other than the group links. There are 31715 users with at least one tag, 98.7% of which posted the same tag with at least one other user. We used the data for a spam filtering task, considering that the sensitive attribute that spammer users would be interested to hide is the binary attribute of whether someone is a spammer or not.

5.2 Experimental setup

We ran experiments for each of the five presented attack models: an attack in the absence of link and group information (baseline model), the link-based friend-aggregate attack and blockmodeling-based attack, the group-based attack in which relevant groups are selected and its simpler version which considers all groups. For the first three models, we ran leave-one-out experiments which assume that complete information is given in the network in order to predict the sensitive-attribute of a user. For the group-based approaches we split the data into test and training by randomly assigning each profile to be private with

⁴At <http://www.kde.cs.uni-kassel.de/ws/rsdc08/>.

a probability $n\%$, i.e., assuming $n\%$ private profiles in the network. Groups were marked as relevant to the classification task either based on maximum size cutoff, maximum entropy cutoff and/or minimum percent of public profiles in the group. We measure accuracy, node coverage and group coverage. Accuracy is the correct classification rate, node coverage is the portion of private profiles for which we can find the sensitive attribute, and group coverage is the portion of groups used for classification. The reported results are the averages over 5 different trials for each set of parameters.

5.3 Sensitive-attribute inference results

5.3.1 Flickr

In the absence of link and group information, our baseline achieved a very low, 27.7%, accuracy due to the fact that there was a large class skew: 27.7% of the users are from the United States. The link-based methods in the presence of complete information performed even worse. Using the majority rule on friends' attribute values, the accuracy was 25.7%. This may be due to the diversity of friends and/or the class skew. Using the blockmodeling-based attack performed even worse, with only 8% accuracy. Clearly, Flickr users do not form friendships based on their country of origin and the country of origin of their friends.

The group results were more promising. We tried a large variety of values for each parameter: percent private profiles in the network, maximum size cutoff, maximum entropy cutoff, and minimum percent of public profiles in the group. Here, we report on the ones that provided more insight in terms of high accuracy and node coverage. Figure 4 (a) shows that naively running the classifier on all group memberships, the prediction accuracy was 63.5%. However choosing the relevant groups based on their size made the accuracy go up to 72.1%. This showed that medium to small-sized groups were more informative. Choosing the relevant groups based on their entropy showed even better results (see Figure 4 (b) with the best cutoff being around 0.5. Using the groups with entropy lower than 0.5 resulted in 83% accuracy. The reported results are for groups with a minimum of 50% public profiles per group. We also experimented with groups with more or less public profiles and the results are presented in Figure 4 (c). When homogeneous groups include at least 90% of public users per group, the accuracy went to 100% but only for a very small set of 1 – 3 users. Including groups, regardless of the percent public profiles per group, yielded a lower accuracy but over a larger set of users. Other advantages of choosing relevant groups based on entropy were that it reduced the group space by 71.2% and that SVM was able to train much faster. The disadvantage is that as we prune groups, some of the users do not belong to any groups, thus the coverage over nodes is lower: 52% of the nodes were predicted with 83.3% accuracy. For privacy purposes, this is a strong result, and it means that groups can help an adversary predict the sensitive attribute for half of the users with private profiles with a high accuracy. Figure 4 (d) shows what would happen if there were more users with private profiles in the network. As expected, when there

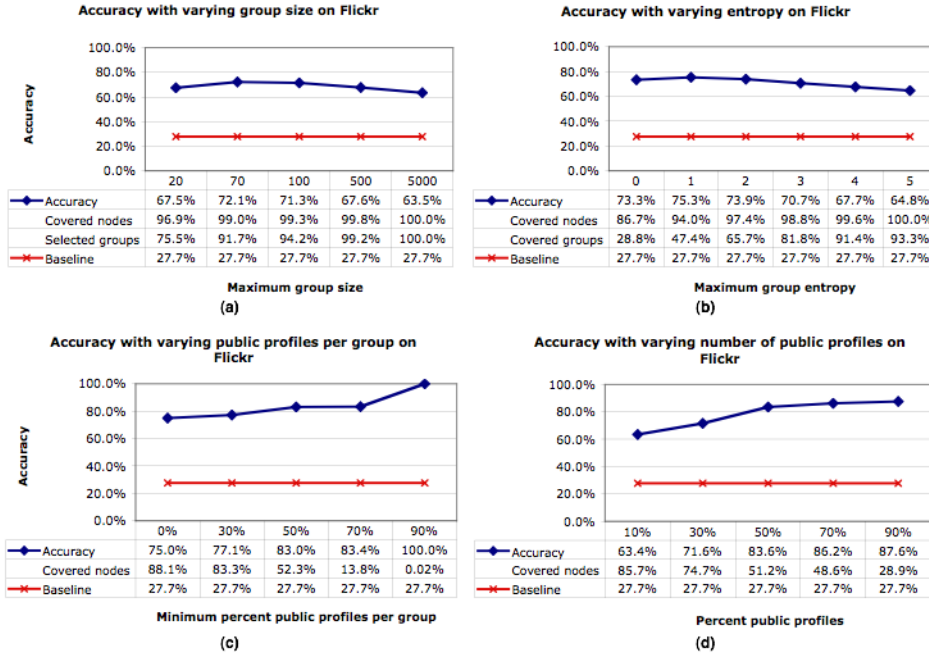


Figure 4: Prediction accuracy on the Flickr dataset when there are 50% private profiles and the relevant groups are chosen based on (a) varying size, (b) varying entropy, and (c) a varying minimum requirement for the number of public profiles per group with a maximum entropy cutoff point at 0.5. Accuracy for various percent of public profiles (d): the more the private profiles in a network, the worse the accuracy and therefore, the better the privacy of users.

are more private profiles, the accuracy is worse and therefore, users' privacy is better preserved. When there are mostly public profiles, the accuracy can go as high up to 84 – 88%. Even in the case of mostly private profiles, the accuracy is relatively high considering the baseline of 27.7%. The reported results are for the case when the minimum portion of public profiles per group is equal to the portion in the overall network and the cutoff for the maximum group entropy is at 0.5.

Looking at the most and least relevant groups provided some interesting insights. The most heterogeneous group, i.e., the one with the highest entropy that our method found is called "worldwidewondering - a travel atlas", and it includes photos and discussion of countries from all over the world. As its name suggests, it pertains to users from different countries and using it to predict someone's country seems useless. Some of the larger homogeneous groups include "Beautiful NC", "Disegni e scritte sui muri" and "Nederland belicht." One of the homogeneous groups has the nondescript name "ponx", and it includes users from the same country. For example, for one user we looked at, member-

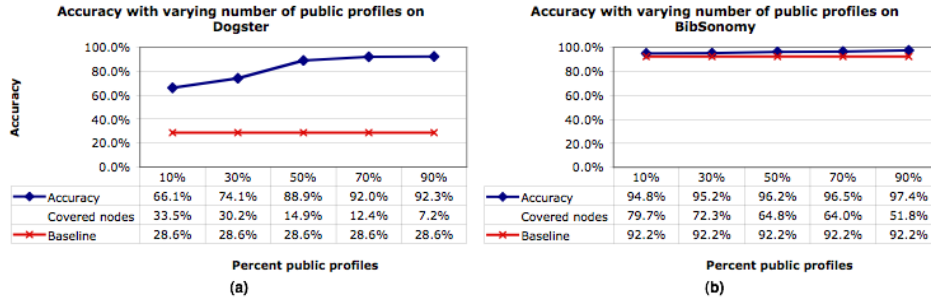


Figure 5: Prediction accuracy on (a) Dogster and (b) BibSonomy.

ships in this group and another homogeneous groups helped us determine that a user who claimed to be from all over the world was most likely from Mexico. The content of his pictures on Flickr confirmed this as well.

5.3.2 Dogster

In Dogster, the baseline accuracy in the absence of groups and links is 28.6%. Including all groups as features led to 66.6% accuracy when there were 50% public profiles, and it had very high coverage over nodes (92.2%). Pruning groups based on entropy led to much higher accuracy (88.9%) but had lower coverage over nodes (14.9%). Figure 5 (a) shows the accuracy and percent covered nodes for various percent private profile assumptions. We tried different options for the maximum group entropy required, and we report on the results for 0.5. The accuracy increased significantly as the number of public profiles in the network increased with one exception: the accuracies for 70% and 90% public profiles did not have a statistically significant difference. A group named "All Fur Fun" was the least homogeneous of all groups, i.e., had the highest group entropy of 2.7. The online profile of the group shows that this is a group that invites all dogs to party together, so it is not surprising that dogs of many different breeds join. The larger homogeneous groups included "Boxers International," "Westies Unite" and "German Shepherds World."

5.3.3 BibSonomy

We used the BibSonomy data to see whether the group-based classification approach can also help in detecting malicious behavior in a social-media website. The binary attribute we were trying to predict was *isSpammer*. There is a large class skew in the data: most of the labeled user profiles are spammer profiles and the baseline accuracy in the absence of links and groups is 92.2%. Using all groups when 50% of the profiles are public to a statistically significant improvement in the accuracy (94.1%) and a very good coverage over users (98.5%), almost all users with tags that at least one other user uses (98.7%). The accuracy results for BibSonomy are presented in Figure 5 (b). We tried

different options for the minimum entropy required, and we report on the results for it being 0. The results suggest that if more profiles are labeled, then more covered spammers would be caught. As in the other results, the coverage gets lower as this happens which in the spam case is actually undesirable. Precision was 99.9 – 100% in all group-based classification cases, meaning that virtually all predicted spammers were such, whereas in the baseline case, it is 92.2%. Some of the homogeneous tags with many taggers include "mortgage" and "refinance."

6 Discussion

From a general data mining perspective, our results suggest that it is possible to predict the attributes of some users with hidden profiles and create better statistics of the attribute's overall distribution. For example, if a marketing company can predict the age and location of users with hidden profiles, it can make its targeted marketing much better. As groups with higher entropy are added, the uncertainty associated with the attribute prediction gets higher, and it gets harder to utilize the existence of diverse groups for node attribute prediction. The results also suggest that it is possible to get valuable data from social media websites, and rather than discarding all the private profiles, it is possible to label some of them with reasonable attribute values and use the data to test data-mining and machine-learning algorithms.

From a privacy perspective, this means that joining more diverse groups preserves privacy better. Therefore, people who are truly concerned about their privacy should consider factors such as homogeneity of the groups they join. Of course, in dynamically-evolving environments, it is harder to assess whether a group will remain diverse as more people join and leave it. Another aspect is the ability to join public groups but display their group memberships only to people they feel comfortable with. For example, social media websites could allow their users to hide their group membership from people who are not their friends. From a data anonymization perspective, our results suggest that a data provider should consider removing groups that are homogeneous in respect to sensitive attributes before publishing a dataset in the public domain.

Other interesting research questions remain to be answered: What are the properties that make a social network vulnerable to a group-based attack? Are profiles on social media websites more or less vulnerable than ones on a purely networking website? What are the specific privacy guidelines that a social network website provider should follow to ensure its users are protected against unintended privacy leaks? Do users with private profiles have group-membership patterns that are different and more privacy-preserving from public-profile members?

7 Conclusion

While having a private profile is a good idea for the privacy-concerned users, their links to other people and affiliations with public groups pose a threat on their privacy. In this work, we showed how one can exploit a social network with mixed profiles to predict the sensitive attributes of users. Using group information, we were able to discover the sensitive attribute values of some users with high accuracy on three real-world social-media datasets.

References

- [1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology*, Nov. 2005.
- [2] E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed-membership stochastic blockmodels. *JMLR*, (in press).
- [3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x: anonymized social networks, hidden patterns, and structural steganography. In *WWW*, 2007.
- [4] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, April 2005.
- [5] M. Hay, G. Miklau, D. Jensen, and D. Towsley. Resisting structural identification in anonymized social networks. In *VLDB*, August 2008.
- [6] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.
- [7] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *TKDE*, 17(4):491–502, April 2005.
- [8] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *ACM SIGMOD*, 2008.
- [9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian. l-diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.
- [10] S. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *JMLR*, 8:935–983, May 2007.
- [11] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets (how to break anonymity of the netflix prize dataset). *SE&P*, 2008.
- [12] M. E. Nergiz and C. Clifton. Thoughts on k-anonymization. In *PDM*, April 2006.
- [13] J. Neville and D. Jensen. Leveraging relational autocorrelation with latent group models. In *ICDM*, 2005.
- [14] P. Sen, G. M. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. Technical Report CS-TR-4905, Univ. of Maryland, 2008.
- [15] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty*, 10(5):571–588, 2002.
- [16] I. Tschantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector learning for interdependent and structured output spaces. *ICML*, 2004.
- [17] Y. Wang and G. Wong. Stochastic blockmodels for directed graphs. *JASA*, 1987.
- [18] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. *PinKDD*, 2007.