To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles

Elena Zheleva
Department of Computer Science
University of Maryland, College Park
elena@cs.umd.edu

Lise Getoor
Department of Computer Science
University of Maryland, College Park
getoor@cs.umd.edu

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 an online 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 practical models that use friendship and group membership information (which is often *not* hidden) to infer sensitive attributes. The key novel idea is that in addition to friendship links, groups can be carriers of significant information. We show that on several well-known social media sites, we can easily and accurately recover the information of private-profile users. 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 with mixed public and private user profiles.

1. INTRODUCTION

In order to address users' privacy concerns, 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. While some users make use of these features, not surprisingly, others are more open to sharing personal information and they disclose more information in their profiles. For example, some people feel comfortable displaying personal attributes such as age, political affiliation or location, while others do not. In addition, most social-media users utilize the social networking services provided 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 surpisingly large amount of information.

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 (this is a commonly occurring scenario in existing social media sites). We define the problem more formally in Section 4. To the best of our knowledge, our work is the first one to look at this problem, and to map it to a relational classification problem in network data with groups.

Here, we propose seven privacy attacks for sensitive attribute inference. The attacks use different classifiers and features, and show different ways in which an adversary can utilize links and groups in predicting private information. We evaluate our proposed models using sample datasets from four well-known social media websites: Flickr, Facebook, Dogster and BibSonomy. All of these websites allow their users to form friendships and participate in groups, and our results show that an attack using the group information achieves significantly better accuracy than the models that ignore it. This suggests that group memberships in social networks have a strong potential for leaking information, and if links and group affiliations are public, users' privacy in social networks is illusionary at best.

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 that in addition to friendship links, group affiliations can be carriers of significant information.
- We show how to reduce the large number of potential groups in order to improve the accuracy of group-based attribute inference.
- We illustrate the privacy implications of publicly affiliating with groups in social networks and discuss how our study affects anonymization of social networks.
- We evaluate our attacks on challenging classification tasks in four social media datasets.
- We show how surprisingly easy it is to infer private information from group membership data.

We motivate the problem in the next section. Then, we describe the data model in Section 3. Section 4 presents the privacy attacks for sensitive attribute inference, and Section 5 provides experimental results using these attacks. Section 6 presents related work, and Section 7 discusses the broader implications of our results.

2. MOTIVATION

Disclosing private information means violating the rights of people to control who can access their private information. Therefore, in order to prevent private information leakage, it is very important to be aware of the ways in which an adversary can attack a social network to learn the private attributes of users. Studies on the challenges of preserving the privacy of individuals in social networks have emerged only in the last few years, and they have concentrated on inferring the identity of nodes based on structural proper-

An earlier version of this paper appears as technical report CS-TR-4922, July 2008.

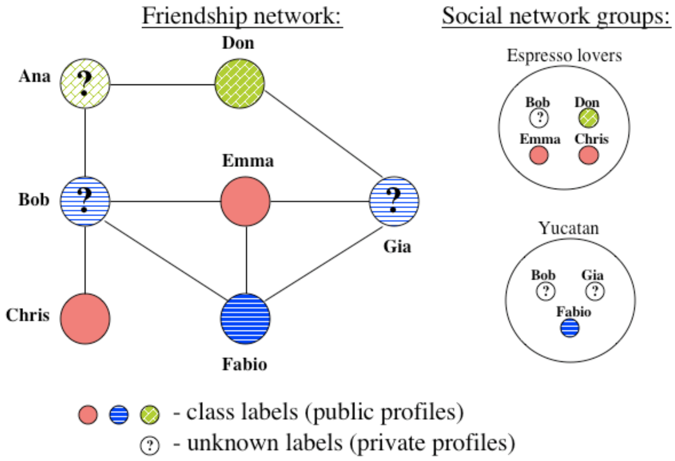


Figure 1: Toy instance of the data model.

ties such as node degree. In contrast, we are interested in inferring sensitive attribute of nodes using some of the approaches developed for relational learning, another active area of research in the last few years.

The novelty of our work is that we study the implications of mixing private and public profiles in a social network. We show that it is very important to be able to make private not only profiles but also friendship links and group memberships. 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 links and even if they do, their friendship links can be found through the backlinks from their public-profile friends. Similarly for group participation information – even if a user makes her profile private, her participation in a public group is shown on the group's membership list. Currently, neither Facebook nor Flickr allow users to hide their group memberships from public groups. It is important that social media website providers protect their users against undesired eavesdropping by informing them of the possible privacy breaches and providing them with the means to be in full control of their private data. Besides undesired eavesdropping by curious people, attacks on private attributes can be used for various purposes by commercial and governmental entities that the user may wish to protect against, including targeted advertising, health care screening, political monitoring, etc.

Our work is also complimentary to work on data anonymization. In data anonymization the goal is to perturb data in such a way that the privacy of individuals is preserved. Even though the goal of our work is not to release anonymized data, it illustrates how data in social networks can be exploited to predict hidden information; this is important in guiding the anonymization process.

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

3. DATA MODEL

We represent a social network as a graph $G = (V, E, H)$, where V is a set of n nodes of the same type, E is a set of edges (the friendship links), and H is a set of groups that nodes can belong to. $e_{i,j} \in E$ represents a directed link from node v_i to node v_j. Our model handles undirected links by representing them as pairs of directed links. 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 also have a set of properties $h.T$.

We assume that each node v has a sensitive attribute $v.a$ 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 $\{a_1 ... a_m\}$. 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. 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 $V_s.A$, the sensitive attributes of the private profiles.

Here, we study the case where nodes have no other attribute information beyond the sensitive attribute. This means that in order to make inferences about the sensitive attribute, we need to use some form of relational classifier. While additional attribute information can be helpful and many relational classifiers can make use of it, in our setting this is not possible because all of the attributes are likely to be hidden in private profiles.

As a running example, we consider the social network presented in Figure 1. It describes a collection of individuals (Ana, Bob, Chris, Don, Emma, Fabio, and Gia), along with their friendship links and information about the interest groups in which they participate. 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 (marked solid), Bob, Gia and Fabio share the same attribute value (marked with stripes), and Ana and Don have a third value (marked with a brick pattern). 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. Thus, this group information can leak information about sensitive attributes, although the manner in which it is leaked is not necessarily straightforward.

4. SENSITIVE-ATTRIBUTE INFERENCE MODELS

The attributes of users who are connected in social networks are often correlated. At the same time, 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 becomes harder to utilize the complex interactions in online social networks for predicting user attributes.

Attribute disclosure occurs when an adversary is able to infer the sensitive attribute of a real-world entity accurately. The sensitive attribute value of an individual can be modeled as a random variable. This random variable's distribution can depend on the overall network's attribute distribution, the friendship network's attribute distribution or the attribute distribution of each group the user joins.

The problem of *sensitive attribute inference* 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 M 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:

$$v_s.\hat{a}_M = \operatorname*{argmax}_{a_i} P_M(v_s.a = a_i; G).$$

where $P_M(v_s.a = a_i; G)$ is the probability that the sensitive attribute value of node $v_s \in V_s$ is a_i according to model M and the observed part of graph G.

We assume that the overall distribution of the sensitive attribute is either known or it can be found using the public profiles. An attack using this distribution is a *baseline attack*. A *successful attack* is one which, given extra knowledge, e.g., friendship links or group affiliations, has a significantly higher accuracy than the baseline attack. The extra knowledge *compromises* the privacy of users if there is an attack which uses it and is successful.

4.1 Attacks without links and groups

In the absence of relationship and group information, the only available information is the overall marginal distribution for the sensitive attribute in the public profiles. So, the simplest model is to use this as the basis for predicting the sensitive attributes of the private profiles. More precisely, according to this model, BASIC, the probability of a sensitive attribute value can be estimated as the fraction of observed users who have that sensitive attribute value:

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

where $|V_o.a_i|$ is the number of public profiles with sensitive attribute value a_i and $|V_o|$ is the total number of public profiles. The adversary using model BASIC picks the most probable attribute value which in this case is the overall mode of the multinomial attribute distribution. In our toy example, the most common observed sensitive attribute is the value that Chris and Emma share. Therefore, the adversary would predict that Ana, Bob and Gia have the same attribute value as well. An obvious 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. Nevertheless, this attack is always at least as good as a random guess, and we use it as a simple baseline. Next, we look at using friendship information for inferring the attribute value.

4.2 Privacy attacks using links

Link-based privacy attacks take advantage of *autocorrela–tion*, the property that the attribute values of linked objects are correlated. An example of autocorrelation is that people who are friends often share common characteristics (as in the proverb "Tell me who your friends are, and I'll tell you who you are"). Figure 2(a) shows a graphical representation of the link-based classification model. There is a random variable associated with each sensitive attribute $v.a$, and the sensitive attributes of linked nodes are correlated. The greying of the other two types of random variables means that the group information is not used in this model.

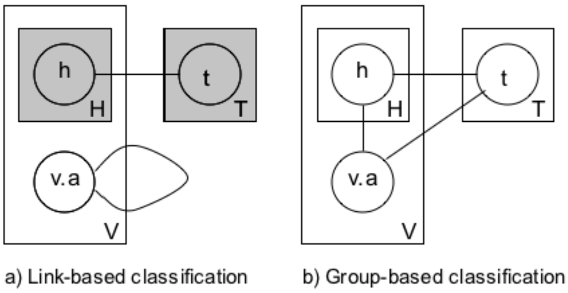


Figure 2: Graphical representation of the models. Grayed areas correspond to variables that are ignored in the model.

4.2.1 *Friend-aggregate model (AGG)*

The nodes and their links produce 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 AGG looks at the sensitive attribute distribution amongst the friends of the person under question. According to this model, the probability of the sensitive attribute value can be estimated by:

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

where $V_o' = \{v_o \in V_o | \exists(v_s, v_o) \in E\}$ and $V_o'.a_i = \{v_o \in V_o' | v_o.a = a_i\}$.

Again, the adversary using this model picks the most probable attribute value (i.e., the mode of the friends' attribute distribution). 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 the one when person's friends are very diverse, as in Gia's case, it will be difficult to make a prediction.

4.2.2 *Collective classification model (CC)*

Collective classification also takes advantage of autocorrelation between linked objects. Unlike more traditional methods, in which each instance is classified independently of the rest, collective classification aims at learning and inferring class labels of linked objects together. In our setting, it makes use of not only the public profiles but also the inferred values for connected private profiles. Collective classification has been an active area of research in the last decade (see Sen et al. [20] for a survey). Some of the approximate inference algorithms proposed include iterative classification (ICA), Gibbs sampling, loopy belief propagation and mean-field relaxation labeling.

For our experiments, we have chosen to use ICA because it is simple, fast and has been shown to perform well on a number of problems [20]. In our setting, ICA first assigns a label to each private profile based on the labels of the friends with public profiles, then it iteratively re-assigns labels considering the labels of both public and private-profile friends. The assignment is based on a local classifier which takes the friends' class labels as features. For example, a simple classifier could assign a label based on the majority of the friends labels. A more sophisticated classifier can be trained using the counts of friends' labels.

4.2.3 Flat-link model (LINK)

Another approach to dealing with links is to "flatten" the data by considering the adjacency matrix of the graph. In this model, each row in the matrix is a user instance. In other words, each user has a list of binary features of the size of the network, and each feature has a value of 1 if the user is friends with the person who corresponds to this feature, and 0 otherwise. The user instance also has a class label which is known if the user's profile is public, and unknown if it is private. The instances with public profiles are the training data which can be fed to any traditional classifier, such as Naïve Bayes, logistic regression or SVM. The learned model can then be applied to predict the private profile labels.

4.2.4 Blockmodeling attack (BLOCK)

The next category of link-based methods we explored are approaches based on blockmodeling [23, 2]. 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. 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_i denote the set of public profiles that have attribute value a_i, and $\lambda_{i,j}$ the probability that a link exists between users in block B_i and users in block B_j. Thus, λ_i is the vector of all link probabilities between block B_i and each block $B_1, ..., B_m$. Similarly, let the probability of a link between a single user v and a block B_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 attack BLOCK is estimated by:

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

where $sim()$ can be any vector similarity function and Z is a normalization factor. We compute maximum similarity using the minimum L2 norm. This model is similar to the class-distribution relational-neighbour classifier described in [16] when the weight of each directed edge is inversely proportional to the size of the class of the receiving node.

4.3 Privacy attacks using groups

In addition to link or friendship information, social networks offer a very rich structure through the group memberships of users. All individuals in a group are bound together by some observed or hidden interest(s) that they share, and individuals often belong to more than one group. Thus, groups offer a broad perspective on a person, and it may be possible to use them for sensitive attribute inference. If a user belongs to only one group (as it is Gia's case in the toy example), then it is straightforward to infer a label using an aggregate, e.g., the mode, of her groupmates' labels, similar to the friend-aggregate model. 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. We propose two models for utilizing the groups in predicting the sensitive attribute – a model which assumes that all groupmates are friends and one which takes groups as classifier features.

4.3.1 Groupmate-link model (CLIQUE)

One can think of groupmates as friends to whom users are implicitly linked. In this model, we assume that each group is a clique of friends, thus creating a friendship link between users who belong to at least one group together. This data representation allows us to apply any of the link-based models that we have already described. The advantage of this model is that it simplifies the problem to a link-based classification problem, which has been studied more thoroughly. One of the disadvantages is that it doesn't account for the strength of the relationship between two people, e.g. number of common groups.

4.3.2 Group-based classification model (GROUP)

Another approach to dealing with groups is to consider each group as a feature in a classifier. While some groups may be useful in inferring the sensitive attribute, a problem in many of the datasets that we encountered was that users were members of a very large number of groups, so identifying which groups are likely to be predictive is important. Ideally, we would like to discard group memberships irrelevant to the classification task. For example, the group "Yucatan" may be relevant for finding where a person is from, but "Espresso lovers" may not be.

To select the relevant groups, one can apply standard feature selection criteria [13]. 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 and 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, a cultural 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: $Entropy(h) = -\sum_{i=1}^{m} p(a_i) \log_2 p(a_i)$ 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. We also consider the confidence in the computed group entropy. One way to measure this is through the percent of public profiles in the group.

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 sensitive attribute value. This step uses only the nodes from the observed set whose 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 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. It shows that there is a dependence between the nodes' sensitive attributes $V.A$, the group memberships H and the group attributes T.

Algorithm 1 Group-based classification model

```
Set of relevant groups $H_{relevant} = \emptyset$
for each group $h \in H$ do
    if $isRelevant(h)$ then
        $H_{relevant} = H_{relevant} \cup \{h\}$
    end if
end for
$trainClassifier(f, V_o, H_{relevant})$
for each sensitive node $v \in V_s$ do
    $v.\hat{a} = f(v.H_{relevant})$
end for
```

5. EXPERIMENTS

We evaluated each of the proposed models to see how effective they were for inferring sensitive attributes in online social networks.

5.1 Data description

For our evaluation, we studied four diverse online communities: the photo-sharing website Flickr, the social network Facebook, Dogster, an online social network for dogs, and the social bookmarking system BibSonomy[1]. For Flickr, the sensitive attribute is the country of the user. For Facebook, the sensitive attributes are gender and political views. For Dogster, the attribute is breed, and for BibSonomy, it is whether or not a user is malicious (a spammer). Table 1 shows important properties of the datasets.

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 API[2] 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 GeoNames[3]. We discarded the profiles with no resolved country location (34%), and also the ones that belonged to a country with less than 10 representatives. The resulting sample contained 9,179 users from 55 countries. There were 47,754 groups with at least 2 members in the sample, and the number of groups of a particular size followed a power-law distribution with many small groups.

Facebook is a social network which allows users to communicate with each other, to form undirected friendship links and participate in groups and events. A part of the Facebook network is available for research purposes [10], and we used it in our experiments. It contains information about all 1,598 profiles of first-year students in a small college. The dataset does not contain social group information but it contains the favorite books, music and movies of the users, and we considered them to be the groups that unify people. 1,225 of the users share at least one group with another person, and 1,576 users have friendship links. All profiles have gender and 965 have self-declared political views. We use six labels of political views - *very liberal or liberal* (545 profiles), *moderate* (210), *conservative or very conservative* (114), *libertarian* (29), *apathetic* (18), and *other* (49).

Dogster is a pet social networking website where dog owners can create profiles describing their dogs, and they post and share information that includes photos and personal characteristics, as well as membership in community groups. Members also maintain links to dog friends and family members. The dataset contains a random sample of 10,000 profiles from Dogster. The dogs that do not participate in any groups were removed from the sample. The remaining 2,632 dogs participate in 1,042 groups with at least two members each, and they have 4,482 links. 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 fourth dataset contains publicly available data from the social bookmarking website BibSonomy[4], in which users can tag bookmarks and publications. Although BibSonomy allows users to form friendships and join groups of interest, the dataset did not contain this information. Therefore, we consider each tag placed by a person to be a group to which a user belongs. 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 affiliations. There are 31,715 users with at least one tag, 98.7% of which posted the same tag with at least one other user. The sensitive attribute is the binary attribute of whether someone is a spammer or not.

5.2 Experimental setup

We ran experiments for each of the presented attack models: 1) the baseline model, an attack in the absence of link and group information (BASIC), 2) the friend-aggregate attack (AGG), 3) the collective classification attack (CC), 4)

[1] At `http://www.flickr.com`, `http://www.facebook.com`, `http://www.dogster.com`, `http://www.bibsonomy.org/`

[2] At `http://code.google.com/apis/`.

[3] At `http://www.geonames.org/export/`.

[4] At `http://www.kde.cs.uni-kassel.de/ws/rsdc08/`.

Table 1: Properties of the four datasets.

Property	Flickr	Facebook	Dogster	BibSonomy
Number of users	9,179	1,598/965	2,632	31,715
Number of links	941,677	86,007/33,597	4,482	N/A
Number of groups	47,754	2,932/2,497	1,042	132,554
Average in-sample degree	142	108/70	1	N/A
Average number of groups per user	162	24/25	1	98
Average group size	31	10/9	3	9
Largest group size	4,527	290/221	118	7,182
Percent links between nodes with the same label	23.5%	49.9%/40.3%	-	N/A
Number of possible labels	55	2/6	7	2
Sensitive attribute	*location*	*gender/polviews*	*breed category*	*spammer*

Table 2: Attack accuracy assuming 50% private profiles. The successful attacks are shown in bold.

Attack model	Flickr	Facebook (gender)	Facebook (polviews)	Dogster	BibSonomy
BASIC	27.7%	50.0%	56.5%	28.6%	92.2%
Random guess	1.8%	50.0%	16.7%	14.3%	50%
BLOCK	8.8%	49.1%	6.1%	-	-
AGG	28.4%	50.2%	57.6%	-	-
CC	28.6%	50.4%	56.3%	-	-
LINK	**56.5%**	**68.6%**	58.1%	-	-
CLIQUE-LINK	**46.3%**	51.8%	57.1%	**60.2%**	-
GROUP	**63.5%**	**73.4%**	45.2%	**65.5%**	**94.0%**
GROUP (50% node coverage)	**83.6%**	**77.2%**	46.6%	**82.0%**	**96.0%**

the flat-link attack (LINK) and 5) the blockmodeling attack (BLOCK), 6) the groupmate-link attack (CLIQUE) and 7) the group-based classification attack (GROUP). For the GROUP model, we present results on both the simpler version which considers all groups and the method in which relevant groups are selected. For the BLOCK model, we present leave-one-out experiments assuming that complete information is given in the network in order to predict the sensitive-attribute of a user. For the AGG, CC, LINK, CLIQUE and GROUP models, we split the data into test and training by randomly assigning each profile to be private with a probability n%. For LINK and GROUP, we used an implementation of SVM for multi-value classification [22].

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. For each experiment, 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 predict the sensitive attribute, and group coverage is the portion of groups used for classification. The reported results are the averages over 5 trials for each set of parameters. We consider an attack to be successful if its average accuracy minus its standard deviation was larger than the baseline accuracy plus its standard deviation.

5.3 Sensitive-attribute inference results

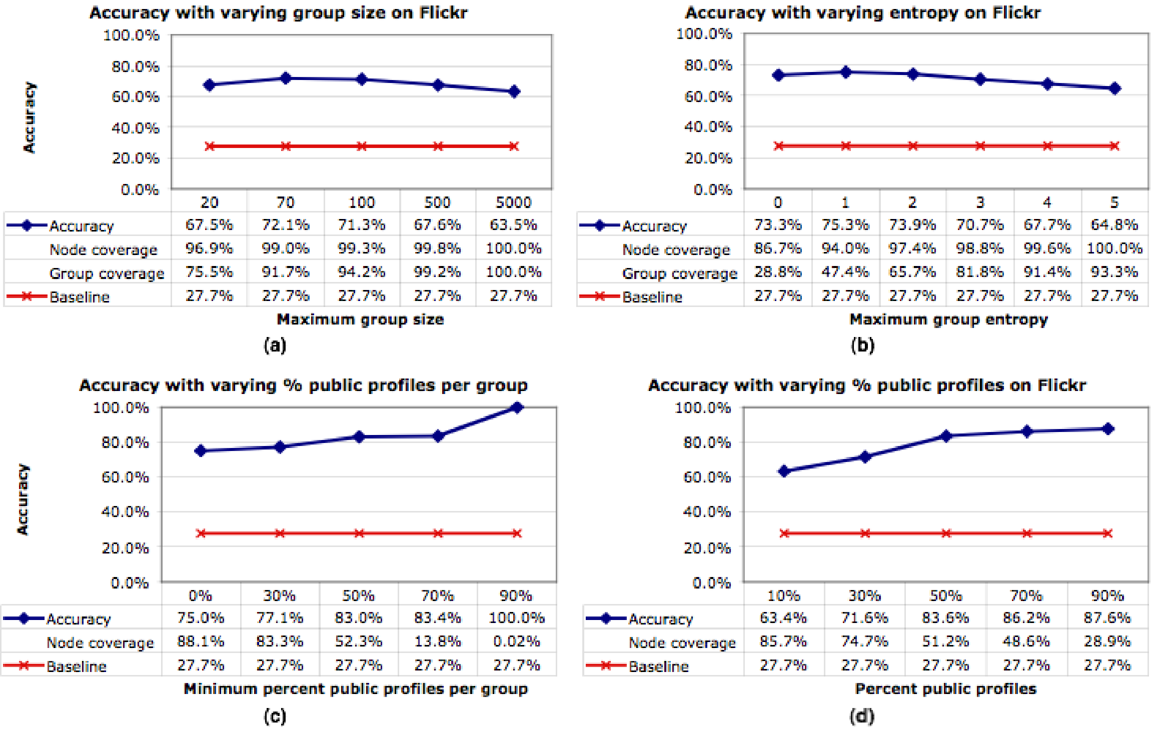
Table 2 provides a summary of the results, assuming 50% private profiles. We see a wide variation in the performance of the different methods. The last line shows the accuracy for half of the users who participate with at least one other user in a group. We also present experiments for varying % of private profiles (Figure 3(d) and Figure 5).

5.3.1 Flickr

Link-based attacks. Not surprisingly, in the absence of link and group information, our baseline achieved a relatively low accuracy (27.7%). However, surprisingly, the link-based methods AGG and CC also performed quite badly. AGG's accuracy was 28.4%, predicting that most users were from the United States. The iterative collective classification attack, CC, performed slightly, but not significantly, better (28.6%). Clearly, Flickr users do not form friendships based on their country of origin and country attribute in Flickr is not autocorrelated (only 23% of the links are between users from the same country). Another possible explanation is that the class had a very skewed distribution which persisted in friendship circles. The blockmodeling attack, BLOCK, performed worse, with only 8.8% accuracy, showing that users from a particular country did not form a natural block to explain their linking patterns. The only successful link-based attack was the ”flattened” link model, LINK. With simple binary features, it achieved an accuracy of 56.5%. We performed experiments based on both inlinks and outlinks, as well as ignoring the direction of the links. The results were slightly better using undirected links, and these are the results we report.

From a privacy perspective, the results from the link-based models are actually positive, showing that in this dataset, exposing the friendship links is not a serious threat to privacy for the studied attribute. The only model which performed well, LINK, shows that if an adversary tries to predict private attributes of users using it, then he has almost a 50-50 chance of being wrong.

Group-based attacks. Next, we evaluate the attacks which used groups. For the CLIQUE model, we converted the groupmate relationships into friendship relationships. This led to an extremely high densification of the network. From an average of 142 friends per user, the average node degree became $7,239$ (out of maximum possible $9,178$). Since the CLIQUE model can use any of the link-based models, we chose to use it with the LINK model because it performed best from the link-based models. This CLIQUE-LINK model has an accuracy of 46.3% and due to the lack of sparsity, its training took much longer time than any of the other approaches.



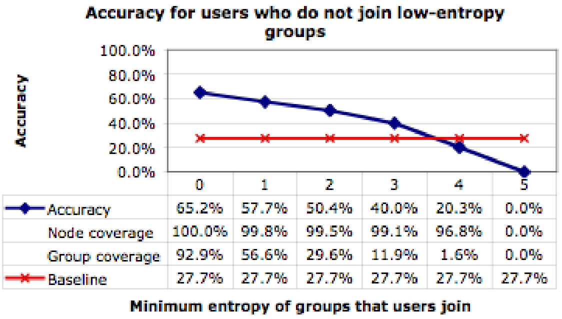
	20	70	100	500	5000
Accuracy	67.5%	72.1%	71.3%	67.6%	63.5%
Node coverage	96.9%	99.0%	99.3%	99.8%	100.0%
Group coverage	75.5%	91.7%	94.2%	99.2%	100.0%
Baseline	27.7%	27.7%	27.7%	27.7%	27.7%

	0	1	2	3	4	5
Accuracy	73.3%	75.3%	73.9%	70.7%	67.7%	64.8%
Node coverage	86.7%	94.0%	97.4%	98.8%	99.6%	100.0%
Group coverage	28.8%	47.4%	65.7%	81.8%	91.4%	93.3%
Baseline	27.7%	27.7%	27.7%	27.7%	27.7%	27.7%

	0%	30%	50%	70%	90%
Accuracy	75.0%	77.1%	83.0%	83.4%	100.0%
Node coverage	88.1%	83.3%	52.3%	13.8%	0.02%
Baseline	27.7%	27.7%	27.7%	27.7%	27.7%

	10%	30%	50%	70%	90%
Accuracy	63.4%	71.6%	83.6%	86.2%	87.6%
Node coverage	85.7%	74.7%	51.2%	48.6%	28.9%
Baseline	27.7%	27.7%	27.7%	27.7%	27.7%

Figure 3: GROUP prediction accuracy on Flickr with 50% private profiles and relevant groups chosen based on (a) varying size, (b) varying entropy, and (c) a varying minimum requirement for the number of public profiles per group (maximum entropy cutoff at 0.5). Accuracy for various percent of public profiles in the network (d): the less public profiles, the worse the accuracy and therefore, the better the privacy of users.



	0	1	2	3	4	5
Accuracy	65.2%	57.7%	50.4%	40.0%	20.3%	0.0%
Node coverage	100.0%	99.8%	99.5%	99.1%	96.8%	0.0%
Group coverage	92.9%	56.6%	29.6%	11.9%	1.6%	0.0%
Baseline	27.7%	27.7%	27.7%	27.7%	27.7%	27.7%

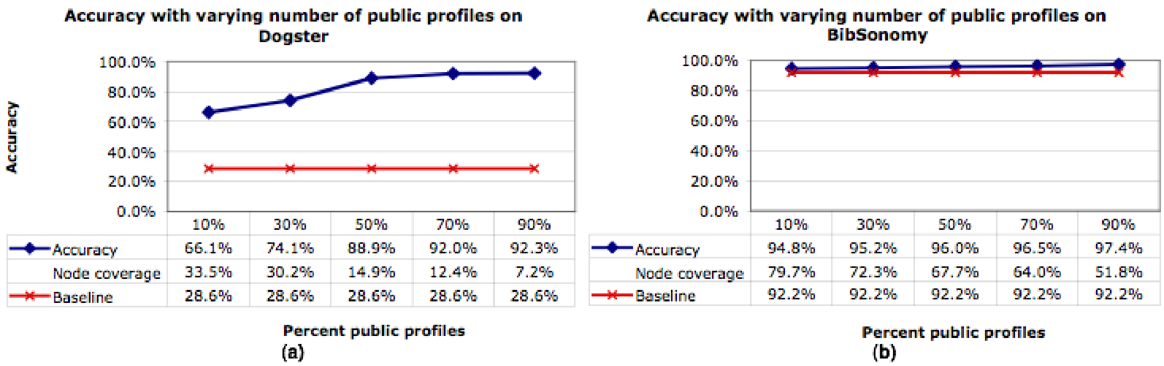
Figure 4: Assuming 50% public profiles, the GROUP accuracy drops significantly if Flickr users with private profiles do not join low-entropy groups.

The group-based classification results were more promising. We evaluated our methods under a wide range of conditions, and we report on the ones that provided more insight in terms of high accuracy and node coverage. Figure 3(a) shows that naïvely running GROUP on all group memberships, the prediction accuracy was 63.5%. However, as larger groups are excluded, the accuracy improves even further (72.1%). This shows that medium to small-sized groups are more informative. Choosing the relevant groups based solely on their entropy shows even better results (Figure 3(b)). Using the groups with entropy lower than 0.5 resulted in the best accuracy. We also pruned groups based on varying percentages of public profiles per group which raised the accuracy even further (Figure 3(c)). Other advantages of choosing relevant groups were that it reduced the group space by 71.2% and that SVM training time was much shorter. The disadvantage is that as we prune groups, some of the users do not belong to any of the chosen groups, thus the node coverage decreases: 51% of the private profile attributes were predicted with 83.6% 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 3(d) shows that as the number of users with private profiles in the network increases, the accuracy gets worse. However, even in the case of mostly private profiles, the GROUP attack is still successful (63.4%). 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 also provides interesting insights. The most heterogeneous group that our method found is "worldwidewondering - a travel atlas." 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*". Other homogeneous groups were related to country but not in such an obvious manner. For example, one of them has the nondescript name "::PONX::" which turned out to be the title of a Mexican magazine. For one user we looked at, this group helped us determine that although he claims to be from all over the world, he is most likely from Mexico.

Insights on privacy preservation. Since including only low-



	10%	30%	50%	70%	90%
Accuracy	66.1%	74.1%	88.9%	92.0%	92.3%
Node coverage	33.5%	30.2%	14.9%	12.4%	7.2%
Baseline	28.6%	28.6%	28.6%	28.6%	28.6%

	10%	30%	50%	70%	90%
Accuracy	94.8%	95.2%	96.0%	96.5%	97.4%
Node coverage	79.7%	72.3%	67.7%	64.0%	51.8%
Baseline	92.2%	92.2%	92.2%	92.2%	92.2%

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

entropy groups significantly boosts the success of the group-based attack, we conjectured that not participating in low-entropy groups helps people preserve their privacy better. Figure 4 shows that if users with private profiles do not join low-entropy groups, then GROUP is no longer successful.

5.3.2 Facebook

We performed the same experiments for Facebook as for Flickr but we omit the figures due to space constraints. We provide a summary of the results here.

Link-based attacks. In predicting gender, we found that while AGG, CC and BLOCK performed similarly to the baseline, LINK's accuracy varied between 65.3% and 73.5%. In predicting the political views, the link-based methods performed similarly to the baseline as Table 2 shows. LINK's average accuracy was not significantly different from the rest. We also performed binary classification to predict whether someone is liberal or not and the results were similar. The best-performing method was LINK with 61.8% accuracy. From privacy perspective, this result means that while it is easy to predict gender, it is hard to predict the political views of Facebook users based on their friendships.

Group-based attacks. The GROUP attack was successful in predicting gender (73.4%) when using all groups. Selecting groups that have at least 50% public profiles per group raised the accuracy by 4% but dropped the node coverage by a half. Predicting political views with GROUP was not successful (45.2%); some possible explanations are that the groups we considered are not real social groups and that books, movies and music taste of first-year college students may not be related to their political views. The relatively low number of groups may also have had an effect.

5.3.3 Dogster

Link-based attacks. Due to the fact that this was a random rather than a snowball sample, there were only 432 nodes with links, and link-based methods are at an unfair disadvantage, so we do not report their results here.

Group-based attacks. In Dogster, the baseline accuracy was 28.6%. CLIQUE-LINK's accuracy was significantly higher (60.2%), as was GROUP's accuracy (65.5%) when there were 50% public profiles. Pruning groups based on entropy led to even higher accuracy (88.9%) but had lower node coverage (14.9%). Figure 5(a) shows the accuracy and node coverage for various private profile percentage assumptions. We tried different options for the maximum group entropy required, and here, 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.

5.3.4 BibSonomy

Group-based attacks. We used the BibSonomy data to see whether the group-based classification approach can also help in predicting whether someone is a spammer or not. 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 leads to a statistically significant improvement in the accuracy (94%) and has a very good node coverage (98.5%); this covers 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 explored different options for the minimum entropy required, and we report on the results for it being 0, i.e., only completely homogeneous groups were chosen. As in the other results, the coverage gets lower when the most homogeneous groups are chosen (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%. The results also suggest that if more profiles were labeled, then more covered spammers would be caught. Some of the homogeneous tags with many taggers include "mortgage" and "refinance."

6. RELATED WORK

To position our work, here, we present a brief overview of related work in privacy and learning in network data.

6.1 Privacy

According to Li et. al. [11], 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. We are interested in attribute disclosure in online social networks using the public profiles, friendship links and group memberships.

The privacy literature recognizes two types of privacy mechanisms: interactive and non-interactive [6]. In the interactive mechanism, an adversary poses queries to a database and the database provider gives noisy answers. In the non-interactive setting, a data provider releases an anonymized version of the database to meet privacy concerns. Even though our work is closer to the non-interactive setting, the goal of our data provider is not to anonymize a dataset but to ensure that users' private data remains private and cannot be inferred using links, groups and public profiles.

Until recently, the literature on anonymization considered only single-table data, in which the rows represent i.i.d. records, and the columns represent record attributes [1, 5, 11, 15, 21]. 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, 8, 14, 17, 18, 24]. For example, in graph data, there is a third type of disclosure attack: *link re-identification* [24]. 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, 8, 14]. It is also possible to link records in anonymized data to external relational data sources to disclose attribute values [17]. 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. It is important to be aware of the different possible privacy attacks in order to guide anonymization techniques.

He et al. [9] provide an interesting study on the use of friendship links in predicting private attributes in a LiveJournal sample in which the friendship links are given. They create synthetic attribute values in the sample, assuming autocorrelation, and show how to use a Bayesian network in predicting sensitive attributes. In contrast, we consider a variety of attacks assuming a richer network structure and we use social groups. We also test the attacks on four networks with real attributes, showing that autocorrelation is not as ubiquitous as expected.

6.2 Learning in network data

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 [7]. Link-based classification in network data, such as social networks, breaks the assumption that data comprises of i.i.d. instances and it can take advantage of autocorrelation, the property that makes the classes of linked objects correlated with each other. For example, political affiliations of friends tend to be similar, students tend to be friends with other students, etc. A comprehensive review of collective classification can be found in the work by Sen et al. [20].

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 [19] 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. Their method assumes that groups do not overlap. Airoldi et al. [2] study mixed-membership 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 relational classification method that makes use of groups with member-set overlaps, and it distinguishes groups that are relevant to classification based on group features such as size and homogeneity.

7. DISCUSSION

Privacy. Our work shows that groups can leak a significant amount of information and not joining homogeneous groups preserves privacy better. People who are truly concerned about their privacy should consider properties of the groups they join, and social network providers should warn their users of the privacy breaches associated with joining groups. 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 privacy aspect is the ability to join public groups but display group memberships only to friends. Currently, neither Facebook nor Flickr allow group memberships to be private and this is a desirable solution to the problem we have discussed.

Surprisingly, link-based methods did not perform as well as we expected. This suggests that breaking privacy in social networks with mixed private and public profiles is not necessarily straightforward, and using friends in classifying people has to be treated with care. We also conjecture that this depends on the dataset. For example, while link-based methods were not very successful in predicting the location of users in Flickr, they may work well in LiveJournal; for example, a study by Liben-Nowell et al. [12] showed that most of the friendship links in LiveJournal are related to geographical proximity. Another important point to consider is the nature of the sensitive attribute we are trying to predict. For example, predicting someone's political views may be a very hard task in general. Recent research by Baldassarri et. al. [4] shows that most Americans are neither consistently liberal nor conservative, and thus labeling a person as one or the other is inappropriate.

It is also important to consider that in some cases, the assumption that unpublished private attributes can be predicted from those made public may not hold. This happens when the attribute distribution in private profiles is very different from the one in public profiles. An extreme example is a disease attribute which shows values for common diseases such as Flu, Fever, etc, in public profiles, whereas more sensitive values such as HIV appear only in private profiles. In a similar example, young people tend to make their age public, and older ones tend to keep it secret. We plan to address this issue in future work.

Data anonymization. Our results suggest that a data provider should consider removing groups that are homogeneous in respect to sensitive attributes before releasing an anonymized dataset in the public domain. All the privacy attacks we studied are also meant to show that more sophisticated anonymization techniques are necessary. The challenge of anonymizing graph data lies in understanding the 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 relational classification problem and we show how to use friendship links, group affiliation and public attribute values as its features to effectively infer private attributes.

Data mining. We show 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 gender 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 becomes harder to utilize the existence of diverse groups for sensitive attribute inference.

Remaining research questions. There are a number of interesting questions that 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?

8. 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 to 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 surprisingly high accuracy on four real-world social-media datasets. We hope that these results will raise the privacy awareness of social media users and will motivate social media websites to enable greater control over release of information and to help their users understand the potential for leaking information.

9. ACKNOWLEDGMENTS

The authors would like to thank Alan Mislove for providing the code for extracting and storing data from Flickr, Jen Neville for a useful discussion on autocorrelation, Michael Hay for valuable feedback on an initial draft of this paper, and Galileo Namata for help in using GIA, a tool with collective classification capabilities. This work was in part supported by NSF under Grants No. 0746930 and No. 0423845.

10. REFERENCES

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Approximation algorithms for k-anonimity. *JPT*, Nov. 2005.

[2] E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed-membership stochastic blockmodels. *JMLR*, 9:1981–2014, 2008.

[3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x: anonymized social networks, hidden patterns, and struct. steganography. In *WWW*, 2007.

[4] D. Baldassarri and A. Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446, September 2008.

[5] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, April 2005.

[6] C. Dwork. Differential privacy. In *ICALP*, 2006.

[7] L. Getoor and B. Taskar, editors. *Introduction to statistical relational learning.* MIT Press, 2007.

[8] M. Hay, G. Miklau, D. Jensen, and D. Towsley. Resisting structural identification in anonymized social networks. In *VLDB*, August 2008.

[9] J. He, W. Chu, and Z. Liu. Inferring privacy information from social networks. In *ISI*, 2006.

[10] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time. *hdl:1902.1/11827*.

[11] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anon. and l-diversity. In *ICDE*, 2007.

[12] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, August 2005.

[13] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *TKDE*, 17(4):491–502, April 2005.

[14] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *SIGMOD*, 2008.

[15] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.

[16] S. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *JMLR*, 8:935–983, May 2007.

[17] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. *S&P*, 2008.

[18] M. E. Nergiz and C. Clifton. Multirelational k-anonymity. In *ICDE*, April 2007.

[19] J. Neville and D. Jensen. Leveraging relational autocorrelation with latent group models. In *ICDM*, 2005.

[20] 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.

[21] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *IJU*, 10(5), 2002.

[22] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector learning for interdependent and structured output spaces. *ICML*, 2004.

[23] Y. Wang and G. Wong. Stochastic blockmodels for directed graphs. *JASA*, 1987.

[24] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. *PinKDD*, 2007.