Information Exposure From Relational Background Knowledge on Social Media

Shuo Liu
Georgetown University
Washington, D.C., USA
sl1539@georgetown.edu

Lisa Singh
Georgetown University
Washington, D.C., USA
lisa.singh@georgetown.edu

Kevin Tian
Georgetown University
Washington, D.C., USA
kt493@georgetown.edu

***Abstract*—While some users share large amounts of information, others share very little. However, even with limited amounts of sharing, users may still have high levels of exposure. Previous research has shown that for certain attributes like gender, adversaries can determine a target's hidden attribute value by taking a majority vote of its community or by finding others in the site population with similar profiles. However, for some attributes, these attacks fail because of the diversity of the attribute value in the community. In this paper, we present a new privacy attack - a *relational background attack* (RBA), where an adversary builds inference models for a hidden attribute of the target by using the target's relational background set. Doing this allows the adversary to build a "biased" model that captures the significant local features for inferring the hidden attribute. We empirically demonstrate the effectiveness of this attack on a special case of the relational background set (a local community) using a Twitter data set. We then consider the case when an adversary only has access to different subsets of the target's local community, and show that the attack can still be conducted effectively with certain approximations of the target's local community.**

***Index Terms*—data privacy, Twitter, information exposure**

I. Introduction

Given the recent exploitation of social media data, data privacy on social media is a growing concern. While some users share significant amounts of personal information, others share very little. However, even with limited amounts of sharing, users may still have high levels of exposure because of available background knowledge. Our goal is to understand the impact of background knowledge on data privacy.

A background knowledge attack occurs when an adversary uses previously learned knowledge to infer sensitive information of a target user. Privacy researchers have shown that background knowledge attacks are especially hard to defend against in the real world [1]. It has also been shown that knowledge about one's community [2] or group membership [3] can be used to reduce a user's data privacy. Typically these studies use a majority vote among community members. However, the user may be a minority in a community with respect to the hidden attribute. While previous work has shown that as the number of possible values for the attribute of interest increases, the accuracy of inferred hidden values of the attribute of interest decreases [4], we will show that this decline does not mean those attributes are not vulnerable.

Specifically, this paper defines and demonstrates a *relational background attack* (RBA), where background knowledge is generated not from a random sample, but from a *relational background set* of the target user. A toy example of the attack is shown in Fig.1. On the left side, we see that an adversary is interested in determining a target user's occupation. The target user has posted some information on a particular social media site, but not his occupation. In order to determine his occupation, the adversary uses the social media site's Application Program Interface (API) to collect the profiles of users in the target user's relational background set, e.g., those users who are part of the target user's community on the social media site, builds an inference model based on these profiles, and infers the hidden information of the target user using this inference model. In the example attack in Fig.1, the majority occupation ("*Writer*") among the community profiles is not the occupation of the target user ("*Doctor*"). Also, not everyone in the relational background set has an occupation since different users share different information on social media.

Given this example, this paper looks to answer the following questions: (1) can meaningful background knowledge be generated from the population of the target individual's communities, (2) can an adversary use this generated background knowledge to determine hidden values of target individuals on social media, and (3) how much information is needed from the relational background set of the target user in order to obtain reasonable inference accuracies? Answering these questions will help to better understand this form of information leakage.

Toward that end, this paper makes the following contributions: 1) we define the relational background attack on Twitter; 2) because APIs may limit the amount of information an adversary can obtain, we present different ways to conduct the attack when an adversary can only obtain partial knowledge, i.e. data for a subset of neighbors; and 3) we execute the attack using Twitter data and show its effectiveness on different hidden attributes using different community sizes, and on users who share different amounts of information.

II. Related literature

Social media privacy continues to be an active area of research. There are a number of papers investigating methods for learning demographic features of users with their social

978-1-7281-8206-3/20/$31.00 ©2020 IEEE
DOI 10.1109/DSAA49011.2020.00041

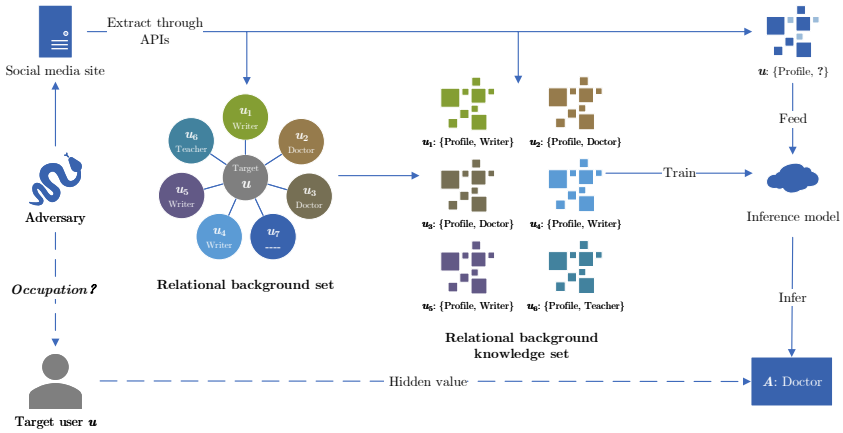


Fig. 1: An example of RBA.

media data, including location [4], age [5], gender [6], and those trying to determine multiple features [2], [7]–[11]. Our paper largely builds off the insights and algorithms developed by Moore et al. [12], where the authors show that using a social media site's population norms can give insight into a target user's hidden attributes. Bergsma and Van Durme [13], Colleoni et al. [14], Rao et al. [6], and Fang et al. [15] explored potential inference algorithms that exploited the unstructured data of a user to predict latent attributes. Chaabane [16], Tang [17], and Al Zamal [7] looked at using supplemental information such as interests and neighbor attributes to train different prediction algorithms. Culotta [18] used an external website's web tracking data to predict the demographic breakdown of social media users. Researchers have also explored algorithms that look at attributes across sites to infer users' information [10], [19]–[21]. A recent thread of research shows that privacy breaches are possible even if social media data are anonymized to share with advertisers or researchers [22]. While the spirit of our work is similar to these previous works, our attack model differs and our experimental design.

A number of papers investigate ways to exploit relationship information among social media users for different attribute inference tasks, where a "relation" may be linkages/connections (e.g. friendship or follower), or same group membership (e.g. hashtags on Twitter or groups on Facebook) [2], [3], [23], [24], [25], [26]. Focusing on the most relevant work, Zheleva and Getoor [3] showed that using the majority vote of an attribute based on groups and linkages (friendships) as features is effective for predicting the values of latent attributes such as gender, political view or geographical location of a target user. While using the neighborhood attribute value is effective for more homogeneous networks, our work extends the ideas to exploit relational information from more heterogeneous neighbors. Mislove et al. [2] explored possible inference methods based on both global and local community detection algorithms to determine a user's hidden value. Tu et al. [27] proposed a method to identify professions by first using traditional classifiers and then refining the results by using connections between target users and different communities based on professions. While both of these works use inference models, they assume that the entire network is available. In this paper, we conduct the relational background attack on both the target user's full network and different subsets of the network, showing its effectiveness under different neighborhood conditions. This is important since APIs do not always give access to all the relationships associated with an individual.

Finally, we consider the variability in the amount of information users choose to share on social media. Those users who rarely or never generate information, instead using social media as a tool for information-consumption, are called passive users. Passive users make up an important fraction of social media. For example, approximately 40% of Facebook [28] and Twitter [29] users are passive users. Previous research suggests that the followees, i.e., people that a user is following, can help infer interests of passive users, leveraging different types of information such as tweets, usernames, biographies [30] and titles of Twitter *lists* [31]. Although this research is conducted from a user-modeling point of view, it still shows that there exists a potential vulnerability with regards to privacy for users, including passive ones, via their followees. Our findings are consistent with previous research. We will show, that the RBA is effective for both passive and active users, and therefore, sharing less does not imply improved privacy.

III. Attack Model

A. Users and relationships

One of the primary goals of social media sites is to connect people. Social media sites accomplish this in different ways. One way is by trying to increase the amount of interaction among users. Each interaction between users can be viewed as a type of relationship. Examples of relationship types between two users include following, mentioning, and liking. Relationship types among multiple users include group membership and participation in the same event. In this paper, we focus our discussion on social connections between users as our

relationship type since they are explicitly identified on social media.

We consider the profile $\mathcal{U}$ of a user u as a tuple, consisting of an unique identifier ID, some number of key-value pairs $\{\mathcal{A}_j\}_{j=1}^{m}$ that belong to the user, and a relational background set $\mathcal{S}$. Because social media data are schemaless (different users share different attributes), we represent our data as a set of key-value pairs $\mathcal{A}_j = (A_j, a_j)$, where A_j is an attribute name, and a_j is its corresponding attribute value. An attribute value can be either a singleton or an unordered set. Examples of key-value pairs are $(\textit{gender}, \{\texttt{male}\})$ and $(\textit{skills}, \{\texttt{Python}, \texttt{C++}\})$.

A user u is a member of a relational background set $\mathcal{S}$, i.e., a set of users on the social media site that are connected to our target u, based on one or more relationship types. An example of a relational background set is a community $\mathcal{C}$, where *community* is a structural property on a network in which nodes are densely connected. Communities can be determined using direct linkage information or using community detection algorithms. Generally, defining a relational background set requires the target user u and the relationship $\mathcal{R}$ to determine the relational background set, $\mathcal{S}(u, \mathcal{R})$.

Formally, we define the user's data as $\mathcal{U} = \big(ID, \{\mathcal{A}_j\}_{j=1}^{m}, \mathcal{S}\big)$. A user might choose to publish a subset of his/her attributes on social media, while hiding the rest. We denote the public profile of user u as $\mathcal{U}^P$, while the hidden component of his profile as $\mathcal{U}^H$. Clearly, $\mathcal{U} = \mathcal{U}^P \cup \mathcal{U}^H$ and $\mathcal{U}^P \cap \mathcal{U}^H = \varnothing$.

B. Relational background attack

In this section, we formally define a new privacy attack, the *relational background attack*. Given an adversary T with knowledge of a user's public profile, $\mathcal{U}^P$, the goal of the adversary T is to recover the value of an attribute of interest A, which is part of user u's hidden profile $\mathcal{U}^H$. Formally, the input of this model is the target user u on a social media site, its public profile $\mathcal{U}^P$, and an attribute of interest A; the output will be a prediction of value a on attribute A for the target user u.

In this model, we assume T is capable of partially acquiring relational information on the social media site u has joined. By *partially* we indicate it might be impossible for the adversary to acquire all the relational information for u on that social media site. However, we assume that T can always collect part of it. Using public relationship information about u, T can then obtain a relational background set $\mathcal{S}$ of the target. Such a set could take the form of a local community, direct neighborhood, and so on. In this paper, we will focus on the local community $\mathcal{C}$ as our relational background set, i.e., $\mathcal{S} := \mathcal{C}$. We also assume that T can partially acquiring public profiles whose value of A can be extracted. In other words, a subset of the community shares values for A as part of their public profiles.

Based on T's knowledge, the adversary could launch a relational background attack. We divide the attack into four parts: (1) acquiring the *relational background set* of the target user, (2) acquiring public profiles and extracting values of the attribute of interest A for these public profiles to build a relational background knowledge dataset, (3) building inference models based on the relational background dataset, and (4) using the learned model to infer an unknown attribute value for A of the target user, i.e., performing the actual attack.

We have already assumed that only a fraction of linkage information can be acquired, and only a fraction of public profiles of the members in the relational background set can be properly labeled. Given this partial data, it is natural for us to also study how much data is enough to make RBA effective. In practice, although social media sites allow T to gather relational information and public profiles through APIs, there usually are restrictions on quantity or rate of data collection. We incorporate these assumptions to make our attack map to reality. Therefore, studying approximations that are variants of the basic attack will help us understand the sensitivity of the inference with regards to the richness of information shared by community members.

IV. Methodology for conducting attacks

The adversary T gathers the desired relational information and public profiles through the a specific social media API. To determine the target user's community on a social media site, T applies a local community detection algorithm on this network, utilizing the linkage information of the sub-network. After determining the relational background set, T collects public profiles of accounts in the relational background set to build a background knowledge dataset. An inference model is built using this dataset and the public profile of u, also acquired from the API, to predict the attribute value of interest.

The pseudocode for RBA is shown in Algorithm 1. The input of the process is the target user u and his public profile $\mathcal{U}$. The attribute of interest also needs to be specified. Once the local community $\mathcal{C}$ of target user u is determined (line 2), the adversary T can build a background knowledge dataset based on this community. After initialization of the local communities $\mathcal{C}$, the list of feature vectors $\mathbf{F}$, and the list of class labels $\mathbf{T}$ (line 3-5), T will collect the public profiles for all the members in $\mathcal{C}$ (line 7), and attempt to extract the value of attribute A as the label (line 8). If the label is extracted, the feature vector and the label will be added to the relational background knowledge dataset (line 10, 11). Inference models using machine learning methods will be trained using this dataset (line 14). Then, given a feature vector based on the public profile of our target user u (line 15) the model will produce the prediction value on the desired attributes in the hidden profile of u (line 16).

A. Local community detection

One kind of the relational background set $\mathcal{S}$ is the target's community $\mathcal{C}$. Although there is no universally accepted definition, generally, a *community* is considered to be a set of nodes in a network where there are more edges connecting the nodes inside the community than edges linking the community and the rest of the network [32]. While numerous community detection algorithms exist, one well known metric used for

Algorithm 1 Relational background attack with local community

```
function LocalCommunityRBA(u, U^P, A)
    C ← GetLocalCommunity(u)
    F ← [ ]
    T ← [ ]
    model ←new Model()
    for u_i in C do
        U_i^P ←GetPublicProfile(u_i)
        label ←ExtractAttributeValue(U_i^P, A)
        if label ≠NULL then
            F.append(GetFeatureVector(U_i^P, A))
            T.append(label)
        end if
    end for
    model.train(F, T)
    f ←GetFeatureVector(U^P, A)
    return model.predict(f)
end function
```

community detection is *modularity* [33]. Modularity is defined as the fraction of edges that fall within a given partition minus the expected fraction of edges if the edges were randomly distributed. A high score will be assigned to a partition with dense connections among nodes within the community and sparse connections with nodes in other communities. Because it is impossible to know the entire structure of the network, using these global metrics is difficult for this scenario. Clauset [34], Luo et al. [35], and Chen et al. [36] proposed different local modularity measures to evaluate a given local community. At a high level, they introduce the notion of a boundary node that connects to nodes outside the community and greedily maximizes local modularity.

Our local community detection algorithm is a variant of [36]. We start with the node of the target user u, and grow the community D greedily, that is, repeatedly adding a node into the current community that maximizes our local modularity measure M' until no node adjacent to the current community can increase M' if added to the community, i.e. $\Delta M' \leq 0$, where

$$\Delta M' = \frac{E_{\text{in}} + e(v, D)}{E_{\text{in}} + E_{\text{out}} + e(v, \overline{D})} - \frac{E_{\text{in}}}{E_{\text{in}} + E_{\text{out}}}$$

for any $v \in \overline{D}$. E_{in} represents the number of edges with both nodes in D, E_{out} represents the number of edges with one node in D and one node in $\overline{D}$, $e(v, D)$ is the number of edges between v and nodes in in D, and $e(v, \overline{D})$ is the number of edges between v nodes in $\overline{D}$.

Fig. 2 shows an example. In these subfigures, the current local community D of target user u consists of green nodes and u itself. E_{in} is the number of green edges, while E_{out} is the number of red edges. Grey nodes are possible candidates that might be added into the local community on the next step. In Fig. 2a, $E_{in} = 11$ and $E_{out} = 5$; $M' = 11/16 \approx 0.69$. In Fig. 2b, the node that would increase M' the most is highlighted using green dashes, $\Delta M' = 0.125$.

V. Approximations for attacks

The adversary T would like to collect all the public relational information from the network and label all public profiles in the target user's community. In this case, we say T is able to launch a *full knowledge* RBA. That is, T can and will grow the full local community of our target user u, and build the background knowledge dataset based on all the public profiles in this community.

Unfortunately for T, in practice it can be difficult to collect all the desired information. There are many reasons for this. First, many social media APIs set data collection limitations. Next, when building the relational background knowledge dataset, some members in the local community might also hide their values for the attribute of interest. Finally, utilizing the data of every member in the community is not efficient since the data collection and community building processes are both time and storage consuming. Given the situation, the attacker is more likely to conduct a *partial knowledge* RBA.

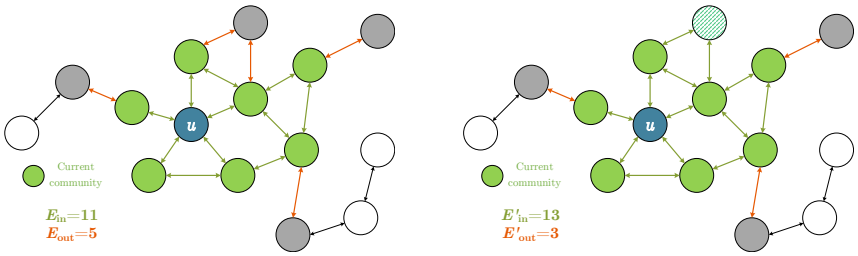
This leads us to consider the following question: how much information is necessary for T to effectively conduct a RBA? To answer this question, we will consider the following approximations of the full local community of our target users: (1) the 1-hop neighborhood of u, (2) a random sample of community members, and (3) a sample based on a fraction of the community members ranked by their centrality metrics. We now go through each of these in more detail.

1-hop neighborhood: The 1-hop neighborhood of a target user u consists of all users directly linked to u. Members inside the 1-hop neighborhood are all closely connected to the target, but they might not be closely connected to each other. A 2016 study shows that Twitter users have an average of 707 followers [37], indicating that the size of 1-hop neighborhoods might be much larger than the size of local communities, possibly resulting in more noise.

Random sampling over community members: One way to reduce the workload of RBA is to use a random sample of the members in the local community. In order to understand the effectiveness of this approximation, we collect the local community, and compare the performance of RBA using the entire community and different random fractions of the community.

Sampling over community by centrality: Centrality characterizes the importance of a node in a network [38]. A higher centrality score indicates more structural importance of a node. For this approximation, we rank nodes in the community using different centrality metrics (degree, closeness, betweenness, and eigenvector) and compare the performance of RBA using community nodes that have the highest centrality values. Here, the assumption is that structurally important members in the community carry more information to build the inference model for determining u's hidden attribute value.[1]

[1]It is possible that structurally important members in the network we obtain may be different from the ones in the actual network. We have designed our experiments to reduce the likelihood of this scenario, but the centrality values should be viewed as part of the approximation since a possibility still exists.



(a) Current community (b) One step after

Fig. 2: Local community metric and local community detection.

VI. Empirical Evaluation

In this section, we present our experimental evaluation. We begin by describing our dataset, followed by the descriptions of the inference models we use to infer values of different attributes. Then we present our experimental results for conducting a RBA using the local community to infer the hidden value of u (partial knowledge attack) and analyze the impact of different approximations of the community. Finally, we simulate a full knowledge attack to understand its effectiveness.

A. Datasets

We use a subset of the dataset collected by Singh et al. [10]. In our sample, there are 257 target users for whom we have their gender, age, country location, and occupation as ground truth attributes. To construct the ground truth, the API of about.me was used, where users self-report their unique identifiers on multiple social media sites. Based on this ground truth, we first collect the follower/following information starting from each target user via the Twitter API, and grow a local community for each target user using the method described in Section IV. We then collect the public profiles of members of each community through the Twitter API. In total, we collected 21,671 public profiles, including our target users' profiles. The average size of the user communities is 84.32, with a maximum of 237 and a minimum of 11. The histogram of the size of communities is shown in Fig. 3. The x-axis is the size of local community we collected, and the y-axis is the frequency of the size. As we can see, the most common sizes of communities are between 10 and 110.

The goal of the RBA is to infer the gender, age, location and occupation of the target users. Table I shows the average percentage of the target's community that shares each *hidden* attribute and the average number of distinct values for each hidden attribute. We see that on average thirty to sixty percent of the communities share each of these hidden attributes. While the fraction is high, the variation is a reminder that people share different information online. The table also shows that the number of distinct values for each community is small, simplifying the inference task. This biasing toward values that are more likely is a major reason this background knowledge attack is successful.

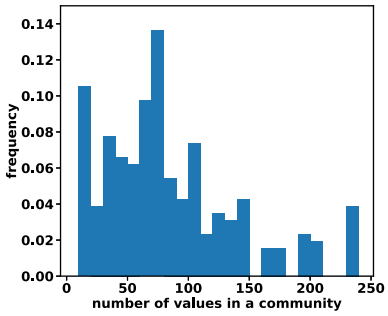


Fig. 3: The size of user communities.

TABLE I: Community statistics

Attribute	Avg. % of community sharing attribute	Avg. domain size per community
location (country)	62.1%	4
gender	40.9%	2
age	52.9%	4
occupation	32.2%	6

B. Experimental setup

In order to simulate the attack, we need to infer values for attributes of the target user's community in order to infer the target's value. We accomplish this by using well established approaches for the community inference (since that is a ground truth value for the community member) and classic models for the target user. All experiments were run 10 times, and their average accuracies are reported. For each hidden attribute, we show the accuracy of determining the value for the target user. We pause to mention that an extensive sensitivity analysis was conducted for every algorithm and every attribute. However, due to space limitations, we only show the best results here.

C. RBA Accuracy

1) Location: The location of community members are determined by extracting the location from the self-reported "location" field and "description" field of their public Twitter profiles. To ensure that they are real locations, we validate them using the Google Map API. The locations may be specified at the city or state level, but we maintain the country level value as the community member's location value.

TABLE II: Inference accuracy of RBA for location.

Model	Attributes used	Country
Naïve Bayes (vector_len= 200)	Tweets	0.895
k-NN ($k = 10$, vector_len= 200)	Tweets	0.911
k-NN ($k = 10$)	"Description" field	0.746
k-NN on tweets ($k = 10$) + Naïve Bayes on description		**0.926**

TABLE III: Inference accuracy of RBA for gender.

Model	Attributes used	Accuracy
k-NN ($k = 3$, num_topic= 50)	"Name" field	0.462
Naïve Bayes (vector_len=100)	Tweets	0.482
k-NN ($k = 10$, vector_len=200)	Tweets	0.642
0.4 k-NN on name ($k = 3$) + 0.6 k-NN on tweets ($k = 10$)		**0.669**

TABLE IV: Inference accuracy of RBA for binned age

Age bin	# of target users	Accuracy
10-19	12	0.667
20-29	80	0.788
30-39	93	0.731
40-49	35	0.714
50-59	26	0.692
60-69	10	0.300
70-79	1	0.000
overall	257	0.720

For inferring the target user's location value, we test Naïve Bayes and k-NN, using different values of k. We use a bag of words model for each tweet and embeddings of those words. Table II show the accuracy results for location. There are a number of interesting findings. First, in general, this is a very successful attack irrespective of the learning algorithm - with success ranging from 75% to 92%. The "description" field of the users is not as good a feature as using the tweet text. This is surprising since the description field tends to contain biographical information about users. However, it is much shorter than the tweet data. The best performance occurs when we use a linear combination of a k-NN classifier ($k = 10$) built using tweets of users, with sentence vectorization using word2vec pre-trained on a Twitter dataset, and a multinomial Naïve Bayes classifier build using the description of users. It is not surprising that combining both pieces of information leads to a better predictive model.

2) Gender: We extract gender using profile images. The model we used is based on a deep neural networks model proposed in [39] and [40]. We put the profile image of users into this model and predict each community member's gender.

Again, we predict the target user's gender using Naïve Bayes and k-NN. We use the following two features: topics generated from the "name" field using LSI [41], and the tweets with bag-of-word embeddings, as we did for location. Table III shows the accuracy results for gender.As we can see, when using the "name" field, a smaller value of $k = 3$ and a medium number of topics (50) leads to better performance. On the other hand, when using topics generated from tweets of the users an even higher accuracy is achieved using k-NN and a larger size of word vectors. This may be a result of inconsistent sharing of the "name" field, the sharing of inaccurate values, the indistinguishably of some names in terms of gender.

The best attack occurs when we use a linear combination of results from a k-NN classifier with topics generated using LSI ($k = 3$, number of topics $= 50$) on the "name" field, and a k-NN classifier ($k = 10$) built using the tweets of users. There are a number of papers showing very high accuracies for inferring gender on Twitter. Our focus is not on the best accuracy possible, but rather conducting a simple version of the attack and showing that the accuracy is still high, 67% in this case.

3) Age: Similar to gender, we use profile images to determine the ages of community members. Again, we use the deep neural networks model proposed in [39] and [40]. We also determine age by searching for birthday self proclamations in users' tweets. We use a dictionary of phrases to search for phrases like "my 20th birthday", and calculate the user's age based on the timestamp of the tweet containing the birth date. If both methods return a valid age, we use the birthday announcement as the final value.

For the age inference task, we create 10-year bins containing ages 10 to 79, the range that all our targets fall into, and we classify the target user into one of them. We use one-versus-rest Support Vector Machines (SVMs) with a radial basis function kernel, and extract the k-top word stems and k-top hashtags as features where $k = 20$. This setting is used by Al Zamal et al. [7] for age inference. Table IV shows the inference accuracy for different age bins. We see that the inference accuracy among younger users is higher. This is not surprising since younger users are more prevalent on social media,[2] and more likely to have connection with their peers.

4) Occupation: The occupation of community members is extracted by building a dictionary of job related words and searching for them in the "description" field of users. We then map these words into one of the 23 major categories specified in the Standard Occupational Classification (SOC)[3] system, a system developed by the US Bureau of Labor Statistics. If there is a word in the description which can be mapped into one of the categories, the category is recorded. The category with the highest count is taken as the occupation label of the community member. Our target users fall into 17 of the 23 categories.

To infer the hidden occupation of the target user, we use SVMs with a radical basis function kernel. The classifier uses word tokens extracted from the tweets of the users and identify word clusters using word2vec pre-trained on a Twitter dataset. This approach was proposed by Preoţiuc-Pietro [42]. Table V shows the inference accuracy on different occupation categories. Overall, the attack is successful approximately 53% of the time. The deviations occur when the size of the

[2]https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/
[3]https://www.bls.gov/soc/2010/home.htm

TABLE V: Inference accuracy of RBA for occupation

Occupation	# target users	Accuracy
Legal	2	1
Healthcare practitioner	3	0.667
Office support	14	0.643
Production	8	0.625
Business	10	0.6
Science	5	0.6
Management	41	0.585
Computer	40	0.575
Arts, sports and media	58	0.517
Education	31	0.516
Personal care	2	0.5
Farming	2	0.5
Sales	19	0.421
Social service	15	0.333
Healthcare support	4	0.25
Construction	1	0
Military	2	0
overall	257	0.533

TABLE VI: Inference accuracy of RBA on different attributes.

Attribute	Accuracy	Majority vote accuracy
location (country level)	**0.926**	0.833
gender	0.669	**0.693**
age	0.720	**0.724**
occupation	**0.533**	0.415

community is very low, meaning that the feature space is not large enough for mapping language to occupations.

Given our inference approach, the obvious question that arises is whether or not a majority vote of the values within the community is sufficient. Table VI compares the results of inferring each hidden attribute to using the majority vote of the community. The results show that the execution of the RBA is more successful for determining location and occupation, has a similar accuracy to the majority vote for age, and has a worse accuracy for gender. This finding seems to indicate that RBA works better on attributes with a larger domain of values, where it is more likely that the target does not belong to the most frequent class.

D. Approximation Results

In this section, we will investigate whether we can obtain similar accuracy results using smaller fractions of the community for the inference of the target user's hidden value, thereby increasing the practicality of the attack. Recall, Table I shows the fraction of the community labeled for different attributes.

1) 1-hop neighborhood: Typically, a 1-hop neighborhood is the easiest relational background set to collect since on most sites it requires only a single API call, and a community detection algorithm is not needed to determine the neighbors. Table VII compares the accuracy of RBAs between the local community and the 1-hop neighborhood approximations. The results show that executing a RBA using only the 1-hop neighborhood is a reasonable approximation of the RBA using a more robust community structure across all of the attributes in the study with accuracy differences of less than 10%. We attribute the larger drop for some of the inference tasks to a higher level of noise in the 1-hop neighborhood compared to the a local community determined using a community detection algorithm.

TABLE VII: RBA approximation - 1-hop neighborhood

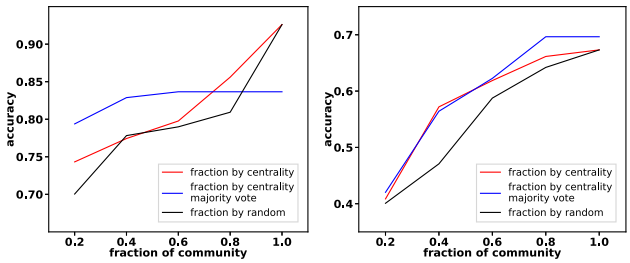
Attribute	RBA accuracy	
	Local community	1-hop neighborhood
location (country level)	0.926	0.852
gender	0.669	0.658
age	0.720	0.704
occupation	0.533	0.492

2) Sample of community: Another way to approximate the local community is to use a fixed fraction of it. One approach is to randomly sample the local community. The other approach is to sample community members based on their importance as measured by different centrality metrics. Fig. 4 shows the accuracy (target user exposure) results. The x-axis shows the fraction of the community used to build the model and the y-axis shows the inference accuracies for the target user. It is no surprise that accuracies increase when using a larger fraction of the community for these different approximations. For all four attributes, sampling using centrality metrics outperforms randomly sampling the community.

We also compared the four different centrality measures mentioned in Section V on country level location. The results are very similar, with a maximum difference of 3% when 80% of the community is used (see Fig. 5). Again, the x-axis shows the size of the sample, and the y-axis shows the accuracy of inference. While all the centrality measure perform well for some of the attributes, we see that closeness centrality outperforms others.

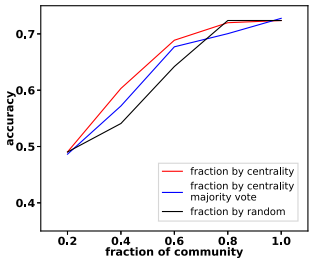
The results of this group of experiments show that for occupation, using labeled public profiles in only 40% of the community can yield a good approximation of the RBA when compared to using data from all the labeled community members. For gender and age, 60% of the local community is needed for a good approximation, and for country location, 80% leads to reasonable results. That is to say, even with a smaller number of public profiles for building the machine learning model, the effectiveness of the RBA is still comparable for these four attributes of the target.

One of our claims in the introduction is that we can maintain a high level of inference even if a user is a passive user as opposed to an active one. Figure 6 shows a histogram of the number of tweets posted by each target user. We see that some target users are not active on Twitter, although most are active. Figure 7 shows the accuracy based on the number of tweets posted by the target user. The figure illustrates that those who are less active have a higher accuracy than those who are more active. This is a reminder of how data exposure is not just in the hands of the information shared by the user. It is also in the hands of the community that the user participates in. We also find that the accuracies drop when the target users share between 1,000 to 10,000 tweets. While this may be caused

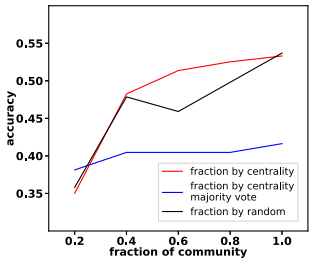


(a) Location (country)

(b) Gender



(c) Age



(d) Occupation

Fig. 4: RBA approximations: samples of local community. The red lines are for a RBA with a sample determined by betweenness centrality, the black lines are for a RBA using a random sample, and the blue lines show the results of using a majority vote for a sample based on centrality measures.

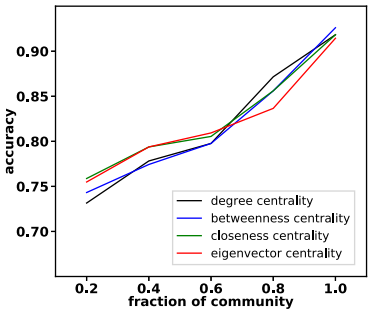


Fig. 5: RBA approximations on location: centrality metrics

by an increase in noise, 6% of the targets share multilingual posts, and a larger subset of them are in this group. Because we use word embeddings, we will be less successful on posts not written in English.

3) Full knowledge RBA: As a reminder, the full knowledge RBA requires every member in the relational background set to be labeled with the hidden attribute value. Given the variability in what people share on social media, it is highly unlikely that everyone in a community will share the target user's hidden attribute. Still, it would be nice to simulate the attack using full knowledge.

Here, we propose a way of approximating the full knowledge RBA:

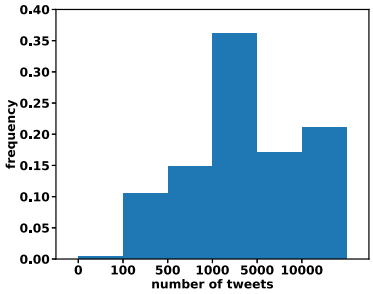


Fig. 6: Tweet frequency for each target user

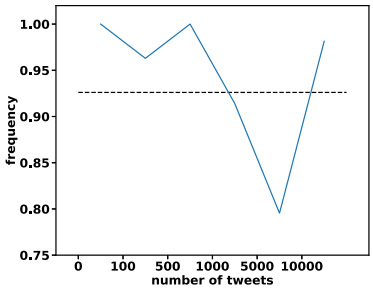


Fig. 7: Accuracy of location inference for passive and active users. The dashed lines indicate the overall accuracies (see Table II).

1) Build an inference model a using all labeled public profiles, and infer labels for those unlabeled public profiles in the local community using model a.
2) Build an inference model b using all public profiles and their labels, whether extracted or inferred, and determine the hidden attribute value of the target user u using model b.

This bootstrapping approximation enables us to label every community member with the attribute we are trying to predict for the target user. Table VIII presents results for this simulated full knowledge attack. We see that the accuracies for the full knowledge attack are similar to those of the partial knowledge attack. This is an indication that inferring the labels of community members lead to more noise than expected. The labels extracted from community members need to be accurate for an effective RBA; therefore, approximating the full knowledge RBA is unlikely to lead to a higher number of successful attacks.

VII. Discussion

The series of experiments simulating RBA using the target user's local community and its approximations show that the relational background set of a target user on social media can be used to effectively build inference models and infer hidden attribute values of the target user. The fact that RBA works at all indicates local patterns do exist and the inference models can capture them. It is important to note that we used basic machine learning models and community detection methods

TABLE VIII: Inference accuracy of approximated full knowledge RBA with local communities on different attributes.

Attribute	RBA accuracy	
	Partial RBA	Approx. full RBA
location (country level)	0.926	0.872
gender	0.669	0.661
age	0.720	0.700
occupation	0.533	0.525

to show the viability of this attack. The effectiveness of the attack may be improved if more sophisticated models are built, e.g. neural networks. The RBA's strength comes from building a "biased" model using a relatively small dataset. Experimental results show that RBA performs better on country level location and occupation. For these attribute, there are more possible values inside each community, meaning that the community members have less homophily. Because of this variability, the hidden attribute values of the neighborhood must be combined with other features in the public profiles of the target's neighborhood in order to build a good inference model. Without this approach to learning, RBA would not be as successful.

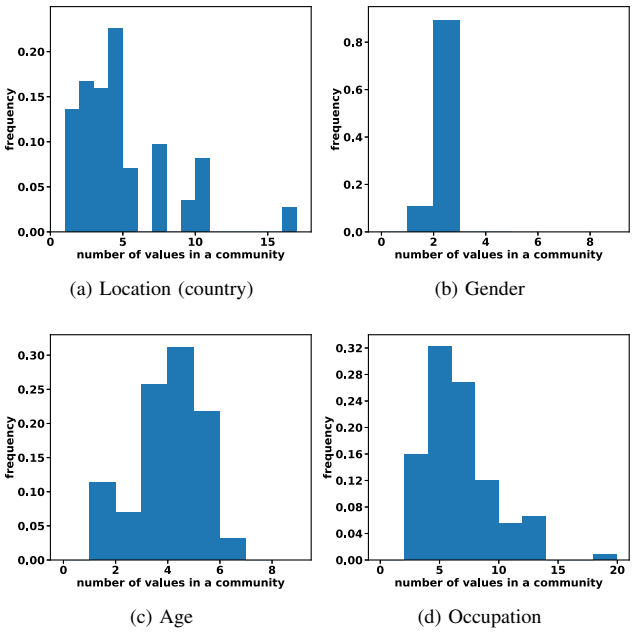
Focusing on the approximation experiments, the results show that the RBA using only the 1-hop neighborhood as an approximation of the local community can achieve an accuracy that is similar to that of a larger fraction of the local community for some attributes, and is less effective for others. This is not surprising since 1-hop neighborhoods are large, and possibly less connected than the local community. Although not all public profiles of members in the relational background set have a value for the attribute in question, RBA remains effective. Experiments that use a smaller fraction of local communities lead to similar accuracy for determining the target's hidden value. That is to say, RBA can be conducted effectively, even with less information.

Finally, we want to identify some of the limitations of the RBA. The RBA does not work well on all attributes, for example, gender and age. The distributions of numbers of possible values on these attributes are shown in Fig. 8b and 8c. Given the fact that those attributes have a smaller domain, and there is more homophily for these attributes values inside communities of our dataset, trying to infer the target value using a biased model might not be a good choice even though local patterns do exist. A simple majority vote seems to give a better accuracy and is less costly to determine. Social media sites are also increasing the limitations of data extraction from their APIs. It might make collective attacks using the RBA unfeasible, since for each target, information from its relational background set is needed. But attacks on single target users are still likely to happen; therefore, it is important to understand the parameters of this attack.

VIII. Conclusion

This paper proposed the RBA, a relational background attack framework that can be used to reveal hidden attribute values of target users on social media by building a biased inference model using only public profiles of members in a relational background set. We studied one particular type of RBA, using the local community as the relational background set. This attack is most effective on attributes whose values are diverse in the community. We also studied different approximations of the RBA using different subsets of the local community. We find that the RBA approximation that used a 1-hop neighborhood, or fractions of the community based on centrality values are good approximations, while the approximated full knowledge RBA introduces noise, limiting its effectiveness.

There are a number of future directions. One is to explore other relational background sets, e.g. using the *groups* on Facebook, where users are related to others but not through bilateral relations, or considering different types of relationships on Twitter. We could then understand the impact of different constructions of the relational background information on the performance of the attack. Another important future direction is to find ways to prevent or reduce the effectiveness of a RBA.



(a) Location (country) (b) Gender

(c) Age (d) Occupation

Fig. 8: Distributions of numbers of possible values in a community for different attributes. The x-axis shows the number of possible values, and the y-axis shows frequencies.

Acknowledgement

This work was supported by the National Science Foundation grant numbers #1934925 and #1934494, and by the Massive Data Institute (MDI) at Georgetown University.

References

[1] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in *International Conference on Data Engineering (ICDE)*, 2006, pp. 24–24.

[2] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks," in *ACM International Conference on Web Search and Data Mining (WSDM)*, 2010, pp. 251–260.

[3] E. Zheleva and L. Getoor, "To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles," in *International Conference on World Wide Web (WWW)*, 2009, pp. 531–540.

[4] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from justin bieber's heart: The dynamics of the location field in user profiles," in *SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 237–246.

[5] R. Dey, C. Tang, K. Ross, and N. Saxena, "Estimating age privacy leakage in online social networks," in *2012 Proceedings IEEE INFOCOM*, 2012, pp. 2836–2840.

[6] D. Rao, M. J. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith, "Hierarchical bayesian models for latent attribute detection in social media." vol. 11, 2011, pp. 598–601.

[7] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors." 2012.

[8] A. Chaabane, G. Acs, and M. A. Kaafar, "You are what you like! information leakage through users' interests," in *Annual Network and Distributed System Security Symposium*, 2012, pp. 1–14.

[9] A. Culotta, N. K. Ravi, and J. Cutler, "Predicting the demographics of twitter users from website traffic data," in *AAAI Conference on Artificial Intelligence*, 2015, pp. 72–78.

[10] L. Singh, G. H. Yang, M. Sherr, A. Hian-Cheong, K. Tian, J. Zhu, and S. Zhang, "Public information exposure detection: Helping users understand their web footprints," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2015, pp. 153–161.

[11] P. Sinha, L. Dey, P. Mitra, and D. Thomas, "A hierarchical clustering algorithm for characterizing social media users," in *Companion Proceedings of the Web Conference 2020*, 2020, p. 353–362.

[12] W. B. Moore, Y. Wei, A. Orshefsky, M. Sherr, L. Singh, and H. Yang, "Understanding site-based inference potential for identifying hidden attributes," in *IEEE International Conference on Social Computing*, 2013, pp. 570–577.

[13] S. Bergsma and B. Van Durme, "Using conceptual class attributes to characterize social media users," in *Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2013, pp. 710–720.

[14] E. Colleoni, A. Rozza, and A. Arvidsson, "Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data," *Journal of Communication*, vol. 64, no. 2, pp. 317–332, 2014.

[15] Q. Fang, J. Sang, C. Xu, and M. S. Hossain, "Relational user attribute inference in social media," *IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 1031–1044, 2015.

[16] A. Chaabane, G. Acs, M. A. Kaafar *et al.*, "You are what you like! information leakage through users' interests," in *Network & Distributed System Security Symposium (NDSS)*, 2012.

[17] C. Tang, K. Ross, N. Saxena, and R. Chen, "What's in a name: a study of names, gender inference, and gender behavior in facebook," in *International Conference on Database Systems for Advanced Applications*. Springer, 2011, pp. 344–356.

[18] A. Culotta, N. R. Kumar, and J. Cutler, "Predicting the demographics of twitter users from website traffic data." in *AAAI Conference on Artificial Intelligence*, 2015, pp. 72–78.

[19] P. Jain, P. Kumaraguru, and A. Joshi, "@i seek 'fb.me': Identifying users across multiple online social networks," in *International Conference on World Wide Web (WWW)*, 2013, pp. 1259–1268.

[20] A. Ramachandran, L. Singh, E. Porter, and F. Nagle, "Exploring re-identification risks in public domains," in *IEEE International Conference on Privacy, Security and Trust*, 2012, pp. 35–42.

[21] J. Ferro, L. Singh, and M. Sherr, "Identifying individual vulnerability based on public data," in *IEEE International Conference on Privacy, Security and Trust*, 2013, pp. 119–126.

[22] G. Beigi and H. Liu, "Identifying novel privacy issues of online users on social media platforms, by ghazaleh beigi and huan liu with martin vesely as coordinator," *SIGWEB Newsl.*, no. Winter, Feb. 2019.

[23] N. Z. Gong and B. Liu, "Attribute inference attacks in online social networks," *ACM Trans. Priv. Secur.*, vol. 21, no. 1, Jan. 2018.

[24] J. Mao, W. Tian, Y. Yang, and J. Liu, "An efficient social attribute inference scheme based on social links and attribute relevance," *IEEE Access*, vol. 7, pp. 153 074–153 085, 2019.

[25] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2016.

[26] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Inferring private information using social network data," in *International Conference on World Wide Web (WWW)*, 2009, pp. 1145–1146.

[27] C. Tu, Z. Liu, H. Luan, and M. Sun, "Prism: Profession identification in social media," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 6, pp. 81:1–81:16, Aug. 2017.

[28] Felim McGrath, "4 in 10 Facebookers now browsing the site passively," Jan. 2015. [Online]. Available: https://blog.globalwebindex.com/chart-of-the-day/4-in-10-facebookers-now-browsing-the-site-passively/

[29] Yoree Koh, "Report: 44% of Twitter Accounts Have Never Sent a Tweet," Apr. 2014. [Online]. Available: https://blogs.wsj.com/digits/2014/04/11/new-data-quantifies-dearth-of-tweeters-on-twitter/

[30] G. Piao and J. G. Breslin, "Inferring user interests for passive users on twitter by leveraging followee biographies," in *Advances in Information Retrieval*, J. M. Jose, C. Hauff, I. S. Altingovde, D. Song, D. Albakour, S. Watt, and J. Tait, Eds. Cham: Springer International Publishing, 2017, pp. 122–133.

[31] ——, "Leveraging followee list memberships for inferring user interests for passive users on twitter," in *ACM Conference on Hypertext and Social Media*, 2017, pp. 155–164.

[32] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[33] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[34] A. Clauset, "Finding local community structure in networks," *Physical review E*, vol. 72, no. 2, p. 026132, 2005.

[35] F. Luo, J. Z. Wang, and E. Promislow, "Exploring local community structures in large networks," *Web Intelligence and Agent Systems: An International Journal*, vol. 6, no. 4, pp. 387–400, 2008.

[36] J. Chen, O. Zaïane, and R. Goebel, "Local community identification in social networks," in *IEEE International Conference on Advances in Social Network Analysis and Mining*, 2009, pp. 237–242.

[37] Ryan MacCarthy, "The average twitter user now has 707 followers," Jun. 2016. [Online]. Available: https://kickfactory.com/blog/average-twitter-followers-updated-2016/

[38] S. P. Borgatti, "Centrality and network flow," *Social Networks*, vol. 27, no. 1, pp. 55 – 71, 2005.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[40] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *IEEE International Conference on Computer Vision Workshops*, 2015, pp. 10–15.

[41] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[42] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras, "An analysis of the user occupational class through twitter content," in *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, vol. 1, 2015, pp. 1754–1764.