A Name-Centric Approach to Gender Inference in Online Social Networks

Cong Tang
Peking University
Beijing, China
tangcong@infosec.pku.edu.cn

Keith Ross
Polytechnic Institute of NYU
Brooklyn, NY 11201
ross@poly.edu

Nitesh Saxena
Polytechnic Institute of NYU
Brooklyn, NY 11201
nsaxena@poly.edu

Ruichuan Chen
MPI-SWS
Kaiserslautern, Germany
rchen@mpi-sws.org

Abstract

Traditionally it has been laborious, via census or otherwise, to obtain a contemporary list of people's names. By crawling Facebook public profile pages of a large and diverse user population in New York City, we create a comprehensive and contemporary first name list, in which each name is annotated with a popularity estimate and a gender probability. We first study the properties of the annotated name list itself, and find that the resulting name popularity follows a Zipf distribution, and that more than 94% of the names can be assigned a specific gender with high probability.

Second, we use the name list as part of a novel and powerful technique for inferring Facebook users' gender. Our name-centric approach to gender prediction partitions the users into two groups, A and B, and is able to accurately predict genders for users belonging to A. Applying our methodology to NYC users in Facebook, we are able to achieve an accuracy of 95.2% for group A consisting of 95.1% of the NYC users. This is a significant improvement over recent results of gender prediction [16], which achieved a maximum accuracy of 77.2% based on users' group affiliations. Moreover, since name is a fundamental attribute of a Facebook user, which can not possibly be hidden from general public (and users also do not intend to use fake names, otherwise it will be hard to locate the user), our gender inference methodology would be difficult to circumvent.

1 Introduction

Current Online Social Networks (OSNs) allow users to control and customize what personal information is available to other users. For example, a Facebook user (Alice) can configure her account in such a way that her friends can see her photos and interests, but the general public can see only her name. In particular, a user has the option of hiding her attributes, such as gender, relationship status, sexual preference, and political affiliation, from the general public.

Alice, of course, knows that Facebook (the company) has full access to any information she has placed on her Facebook pages, including information that she limits to her closest friends and family. However, Alice probably assumes that if she makes available only her name to the general public, third parties have access only to her name and nothing more. Unfortunately for Alice, third parties, by crawling OSNs and applying statistical and machine learning techniques, can potentially infer information – such as gender, age, relationship status, and political affiliation – that Alice has not explicitly made public[16]. To the extent this is possible, third parties not only could use the resulting information for online stalking and targeted advertising, but could also sell it to others with unknown nefarious intentions. This information may also be useful to Facebook itself, e.g., to provide a personalized service to its users, and to understand user characteristics and behaviors. As an example, Facebook has recently requested its users to specify their genders so that it can use proper grammar in features such as news and updates [3].

In this paper, we are concerned with the problem of gender inference for Facebook users. Prior work has considered this problem in the context of Facebook and other OSNs [16]. Their approach is based on a general observation that it may be possible to infer private information about Alice by exploiting information provided by Alice's friends or based on Alice's affiliations with various Facebook groups (public information). For example, if the majority of Alice's friends reveal that they are in their twenties and are Republicans, then it is highly probable that Alice is also in her twenties and is a Republican. Or if Alice is a member of a girls' high school, then she is likely a female. For predicting gender, different inference models based on machine learning techniques were studied in [16]. However, this work only had limited success at gender prediction, with a maximum accuracy

of 77.2% based on users' group affiliations. Moreover, and perhaps more importantly, this method of predicting gender can be circumvented by hiding group affiliations from public profiles, as also mentioned in [16].

Our approach to gender inference is based on users' first names. Our observation is that since name is a fundamental attribute of a Facebook user, which can not possibly be hidden from general public (and users also do not intent to use fake names, otherwise it will be hard to locate the user), a name-centric approach to gender inference would be difficult to evade. To develop such an approach, it is necessary to first analyze users' names.

Our Contributions: We make two-fold contributions:

• *Facebook-Generated Name List:* By crawling Facebook public profile pages for 1.67 million users in New York City, we create a comprehensive and contemporary name list, in which each name is annotated with a popularity estimate and a gender probability. Note that traditionally it has been laborious, via census or otherwise, to obtain a contemporary list of people's names. We study the properties of this annotated name list. After combining nicknames with their "canonical names," we find that the resulting name popularity has a Zipf distribution, and that more than 94% of the names can be assigned a specific gender with high probability.

• *Name-Centric Gender Inference:* The basic properties of the name list suggest a new and powerful technique for inferring gender for users who do not explicitly specify their gender. (We find that 47% of NYC users chose not to disclose their gender, confirming that gender can be regarded as a private attribute for a Facebook user.) In addition to violating users' privacy, such an involuntary gender disclosure could be exploited by marketing companies to launch targeted gender-specific advertising and spamming. For example, a cosmetic company might only be interested in marketing their products to females and can benefit from an automated gender inference method, given that they have access to the names and contact information of a large number of users.

Our name-centric approach to gender prediction partitions the users into two groups, A and B, and is able to accurately predict gender for users belonging to A. Applying our methodology to NYC users in Facebook, we are able to achieve an accuracy of 95.2% for group A consisting of 95.1% of the NYC users. This is a significant improvement over recent results of gender prediction in [16], which achieved a maximum accuracy of 77.2% based on users' group affiliations.

Organization: This paper is organized as follows. Section 2 discusses relevant prior work. We provide our data gathering mechanism and properties of the dataset in Section 3. Section 4 presents the name list generation method and analysis of the properties of the name list. Next, we present our gender predictors and their validation in Sections 5 and 6. Then, we discuss our experiments and results of our gender predictors in Section 7. Finally, Section 8 summarizes our conclusions.

2 Related Work

We review prior work most closely related to the theme of our paper. Most of the prior work is concerned with the problem of inference of one or more private attributes, which is related to our second contribution in this paper. We are, however, not aware of any prior research that analyzes and builds on users' names over OSNs (our first contribution).

Krishnamurthy and Wills examine popular OSNs from a viewpoint of characterizing potential privacy leakage [9]. They also investigated the leakage of personally identifiable information via online social networks, identified different ways in which such leakage occurs, and discussed measures to prevent it [10].

Zheleva and Getoor [16] proposed techniques to predict the private attributes of users in four real-world datasets (including Facebook) using general relational classification and group-based classification. In addition to gender inference (which is the focus of our work), they also looked at prediction of political views. Their accuracy for gender inference with their Facebook dataset, however, is only 77.2% based on users' group affiliations, and the sample dataset used in their study is quite small (1,598 users in Facebook). Moreover, their inference methods can be prevented by hiding group affiliations from public profiles, as mentioned in [16]. In contrast, our inference methodology – based on users' names – would be difficult to circumvent, and we demonstrate its validity on a much larger dataset and achieve much better accuracies.

Other papers [8, 15, 11, 7] have also attempted to infer private information inside social networks. Methods they used are mainly based on link-based traditional Naive Bayes classifiers. However, none of them used name-list to infer users' genders, and we achieve much better accuracies compared to these methods for gender inference.

Jernigan and Mistree [5] demonstrated a method for accurately predicting the sexual orientation of Facebook users by analyzing friendship associations. In particular, they have been successful at predicting whether a Facebook user might be homosexual by correlating similar information provided by user's friends.

Most recently, Mislove et. al. [13] proposed a method of inferring user attributes by detecting communities in social networks, based on the finding that users with common attributes form dense communities. However, people with same attributes, such as gender and birthday,

may not form communities, and thus these attributes may not be accurately predicted using this approach.

3 Crawling and Data Gathering

We develop, in Python, a crawler that visits Facebook user profile pages and stores these pages in a file system. First we collect $90,000$ Facebook user IDs from NYC ("New York, NY" network) as seeds, by using Facebook's feature of browsing users in the same network [1] (visiting `http://www.facebook.com/b.php` with specific network id).

For each seed, we visit each of its friends, then each of its friends' friends, and so on, until we obtain all NYC users reachable from the seed. Because of size of Facebook's social network, the crawler was restricted to profiles only inside NYC. The crawler obtained the profile of pages for 1.67 million users. At the time of the crawl, there were approximately 2 million NYC users. We suspect that most of the non-crawled users are bogus users (see below). Therefore, we crawled nearly all the Facebook users in NYC.

The crawler stores the friends' user IDs in the database. At the time of the crawl, a users profile page is, by Facebooks default privacy setting, public to other users who are in the same network. We then use Facebook accounts in NYC network to download all the profile pages using the user ID obtained. To avoid being banned, we carefully limited the crawling speed.

Eliminating Bogus Users: Although many Facebook users have hundreds of friends and 50% of users visit the site daily (as discussed in [1]), many of the users may be *bogus or dormant*: users who signed up, created a few friends, and disappeared quickly. It may be difficult to predict anything about such users. In order to prevent these bogus users from skewing the results of our study, we remove, from our dataset, the users with less than 5 friends across Facebook.

The size of our compressed dataset is $1,282,563$. Out of the $679,351$ users who specified their genders, the percentage of males is 52.97%. Table 1 shows the properties of the dataset before and after the elimination of bogus users. In this paper, we do all processing on the reduced data set after elimination of bogus users.

4 Using Facebook to Generate an Annotated Name List

We demonstrate that the Facebook network can be used to generate an up-to-date list of first names of the users. In our name list, each first name is annotated with the number of users having this name, the number of male users who have identified themselves with this name, and the number of female users who have identified themselves with this name. To guide the design of our gender inference scheme (as we will discuss in Section 5), we have carefully studied the properties of this list. Our name list and its properties are also of independent interest for other applications, including naming newly born babies and studying naming trends.

Table 1: Properties of the dataset from NYC before and after elimination of bogus users

Property name	Before	After
# users in NYC	$1,668,602$	$1,282,563$
# users who specified gender	$864,543$	$679,351$
% users who specified gender	51.81	52.97
# users who identified as males	$456,591$	$349,730$
# users who identified as females	$407,952$	$329,621$

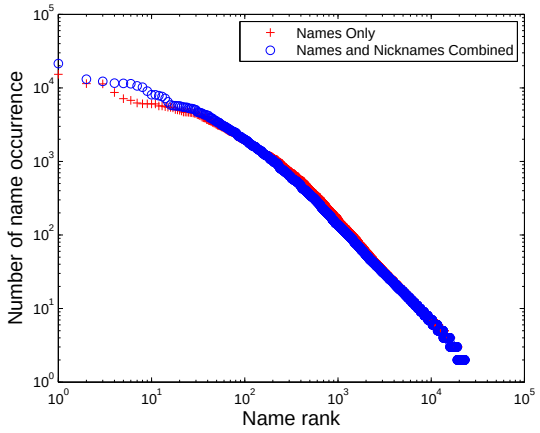

Figure 1: Distribution of Occurrence of Names

We first extract the first names for each of the 1.28 million users and create a crude annotated name list. We then process the crude list to remove entries that are not really names. We remove all one-letter names, all names without a vowel, and names that have been referenced only once. Notice that, for the gender inference analysis in Section 7, we still infer the gender of users whose names have been removed from the list.

After this pre-processing, we obtain a list having $23,405$ names. For each name in the list, we determine the number of users having this name, the number of times it is labeled as male, and the number of times it is labeled as female. We provide this name list online, publicly available at: `http://sites.google.com/site/facebooknamelist/`.

4.1 Combining Names with their Nicknames

As one would expect, we found that many Facebook users identify themselves by using nicknames as their first names. The nicknames, however, might behave as noisy data in our analysis. To avoid this, we design a method that combines nicknames with their "canonical names".

We first create a nickname list which contains 535 nicknames based on resources available on the Internet[2]. For each nickname, we list its canonical names. For example, *Dave*'s canonical name is *David*, and *Stan*'s canonical names are *Stanford* and *Stanley*. Next, we combine the frequency of occurrence of each nickname with frequency of occurrence of its respective "canonical names". Specifically, if a nickname only has one "canonical name", we simply add its frequency of occurrence with the frequency of occurrence of its "canonical name"; if a nickname has multiple canonical names, we calculate its frequency of occurrence based on the frequency of occurrence of each of its "canonical names". For example, let x, y and z be the frequency of occurrence of *Stanford*, *Stanley* and *Stan*, respectively. When combining *Stan* with *Stanford* and *Stanley*, we redefine $x = x + z * [x/(x+y)]$, and $y = y + z * [y/(x+y)]$. After combining nicknames with names, we obtain a name list consisting of $22,993$ entries.

4.2 Analysis of Annotated Name List

Our annotated name list is large and comprehensive (reflecting the broad and diverse demographics of NYC); moreover, this name list is annotated with the number of declared males and females corresponding to each name.

Note that there is a government online service [4] that provides a list of the most popular names for a particular year of birth in the US. However, our annotated name list contains information about NYC Facebook users born both in and outside the US. Moreover, from the public online service, one can only get at most top 1000 names for each year, from which we can obtain a total of $1,736$ male names and $2,023$ female names. Since our name list consists of $22,993$ entries, it is much larger and more diverse than the name list we get publicly from [4].

We now study several interesting properties of this name list.

- **Popularity of names:** Figure 1 shows the distribution of names' occurrence frequency, which roughly follows the power-law distribution with a Zipf parameter $\alpha = 1.3$. We get a more flat Zipf curve after the name/nickname combination. Interestingly, after the combination, there is a single most popular name – *Michael* – which occurs more than $20,000$ times in our name list; then, the next 7 most popular names – *David*, *Elizabeth*, *Jennifer*, *Robert*, *John*, *Joseph* and *Daniel* – occur more than $10,000$ times each. Indeed, these popular names are classic and common American names.

  From Figure 2, we investigate the distribution of names from another perspective. We find that around 18% of names occur only twice. (Note that, when generating the name list, we have removed names that are referenced only once.) Moreover, 80% and 90% of names occur no more than 20 and 50 times, respectively, in our name list.

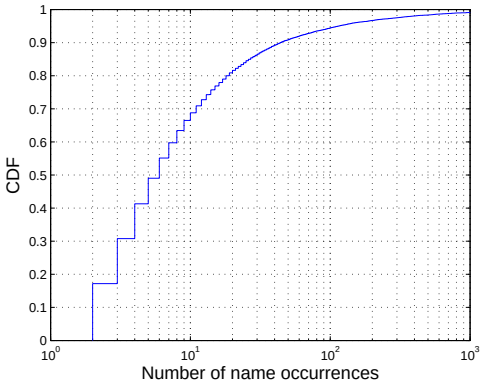

Figure 2: CDF of names' occurrence frequency

- **Gender consistency of names:** Due to various reasons, e.g., the cross-gender names and possible mislabeling, some names may have been labeled as both male and female. Specifically, for each name in our name list, let N_m be the number of users who indicate they are male, and N_f be the number of users who indicate they are female. The fraction of times that a specific name is labeled as male is $f_m = N_m/(N_m + N_f)$. From Figure 3, it is clear that most names are associated with a specific gender; only about 6% of names are ambiguously labeled (i.e., $f_m = 0.5$). This observation will play a central role in our gender inference methodology (as we will discuss in Section 5).

The above analysis of our annotated name list provides some useful insights for gender inference. However, a methodology solely based on the name list will clearly have some difficulties in predicting two types of names: names that have never been labeled and names that are used for both genders. For these two types of names, we have no choice but to resort to other inference methods. In particular, we adopt machine learning techniques (as we will discuss later) to predict these unlabeled and ambiguous names.

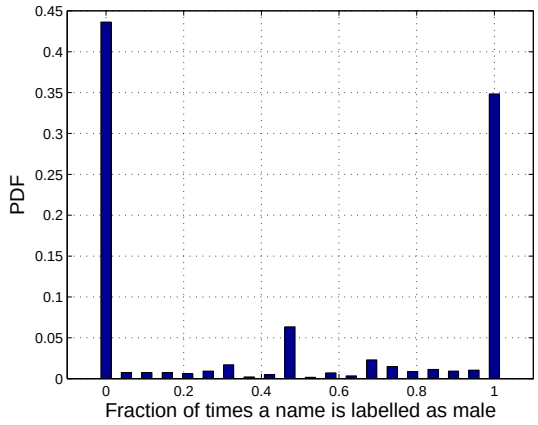

Figure 3: PDF of fraction of times a name is labeled as male

5 Design of Gender Predictors

In this section, we discuss our gender inference methodology and underlying models. In particular, we propose eight predictors for gender inference. These predictors use different features and algorithms, and use different methods of gender inference. We first investigate gender inference using the offline name list and our Facebook generated name list (as discussed in Section 4). We then adopt machine learning algorithms to classify users based on gender. In the classification algorithms, we first use each user's local information from their Facebook profile, such as the user's relationship status, sexual preference and what he/she is looking for. We then use the user's friends' information, such as the fraction of male friends. Finally, we combine our annotated name list predictor with these classification algorithms.

5.1 Offline Name List Predictor (OFL)

We created a first-name list using USA baby name list [4], which consists of $1,736$ male names and $2,023$ female names. Some names in the list, such as "Chris", can be both a male's as well a female's name (e.g., Christopher and Christine, respectively). Such ambiguous names may decrease the gender prediction accuracy, and thus we remove names that are labeled as both male and female from the list. After that, we obtain $1,520$ male names and $1,807$ female names. We then compare each NYC Facebook user to the name list: if user's name can be found in the list, we predict its gender accordingly; otherwise, we only make a random guess to predict the gender.

Table 2: Specification of features used in classification algorithms

No.	**Feature name**	**Value**
1	Relation status: single	0,1
2	Relation status: in a relation	0,1
3	Relation status: engaged	0,1
4	Relation status: married	0,1
5	Relation status: it's complicated	0,1
6	Relation status: in a open relationship	0,1
7	Interested in: men	0,1
8	Interested in: women	0,1
9	Looking for: friendship	0,1
10	Looking for: dating	0,1
11	Looking for: relationship	0,1
12	Looking for: networking	0,1
13	# times the name is labeled as male	0,...,n
14	# times the name is labeled as female	0,...,n
15	Fraction of male friends	$\in [0,1]$

5.2 Facebook Generated Name List Predictor (FB)

Our annotated name list (discussed in Section 4) is much larger and more comprehensive than the aforementioned offline name list. We compare the two lists and find many unpopular first names in our annotated name list that have not been listed in the offline name list. We use Facebook generated name list to predict user's gender. We assign probability to each name in the list based on the fraction of times a specific name is labeled as male, i.e., $f_m = N_m/(N_m + N_f)$. For example, if a name "Tom" has been labeled 95 times as male and 5 times as female, "Tom" is predicted to be a male with probability 95%. We randomly guess for names that do not appear in the list.

5.3 Local Information Predictor (LCL)

Generally, additional information available from user's Facebook profiles, such as relationship status and sexual preferences, can be helpful to our prediction methodology. We select 12 features of a user from his/her profile page (these features are the first 12 features listed in Table 2. Each (binary) feature has a value of 1 if the user corresponds to this feature, and 0 otherwise. For example if the feature "Relation status: single" is 1, the user has indicated he/she is single. We then build our feature vector for a classifier using these twelve features. We choose training data from the profiles of users who have identified their genders, and feed the feature vectors to traditional classifiers.

5.4 Friend Information Predictor (FRND)

In this predictor, we take each user's friends' information into account. We introduce a new feature which is the fraction of a user's male friends (No.15 in Table 2). For the friends who have not specified their gender, we pre-assign genders to them using FB predictor.

5.5 Hybrid Predictors

5.5.1 Name List and Local Information Predictor (FB-LCL)

Now, we combine the FB predictor and the LCL predictor to obtain the FB-LCL predictor. This predictor uses a feature vector for the classifier using the 12 features from the LCL predictor and 2 extra features obtained from the Facebook generated name list: number of times the name is labeled as male, and number of times it is labeled as female (No. 13 and 14 in Table 2).

5.5.2 Name List and Friend Information Predictor (FB-FRND)

In this predictor, we combine the two aforementioned features obtained from the Facebook generated name list and the feature used in FRND predictor into the feature vector for FB-FRND predictor.

5.5.3 Name List, Local and Friend Information Predictor (FB-LCL-FRND)

Finally, we develop a methodology that uses name list, user's local information, and user's friends information. We combine the FB-LCL predictor and the FRND predictor into a single predictor: FB-LCL-FRND. This predictor extends the FB-LCL predictor's feature vector with features used in the FRND predictor.

6 Evaluation of Gender Predictors

6.1 Experimental Setup

We ran experiments for each of the seven predictors (discussed in Section 5). For the LCL, FB-LCL, FB-FRND and FB-LCL-FRND predictors, we choose users who have specified their genders from our data set, generate corresponding feature vectors for each predictor, then split the feature vectors into test set and training set by randomly marking each user's gender as unknown with a given probability. In the following experiments, we use a probability of 50%. We use the Weka toolkit [6] to build classifiers for all of the above training sets. We explored a variety of classifier types and selected Multinomial Naive Bayes (MNB) [12] which yielded the best overall performance in preliminary tests using the training set. In FRND predictor, instead of using MNB classifier, we use a decision tree based classifier J48 [14].

Table 3: Inference accuracy (assuming 50% genders were unknown)

Predictor	Accuracy
OFL	75.5%
FB	92.6%
LCL	66.9%
FRND	60.0%
FB-LCL	94.8%
FB-FRND	94.1%
FB-LCL-FRND	94.6%

6.2 Effectiveness of Gender Predictors

We outline our inference results as follows. A summary of our results is also presented in Table 3.

- The results show that the OFL predictor achieves an accuracy of 75.5% by using the offline name list, in which 55.2% of users' names can be found.
- Our Facebook generated name list significantly improves the inference accuracy to 92.6%, in which 91.7% of users' names can be found.
- Introducing users' local information by using the FB-LCL predictor provides a small gain, increasing the accuracy to 94.8%.
- Introducing friends' information by using the FB-FRND predictor also provides a small gain, increasing the accuracy to 94.1%.
- Friends' information does not provide any additional gain when using the FB-LCL-FRND predictor, because there is some noise along with the friends' information that decreases the prediction accuracy.

6.2.1 Impact of Features in the LCL predictor

We run experiments to determine the local features (attributes) which are most important and useful for gender inference. Specifically, we test four different feature vectors outlined as follows:

1. *Feature Vector 1* is composed of 6 relation status features (No. 1-6 in Table 2) of the user whose gender is to be predicted.
2. *Feature Vector 2* is composed of 2 sexual preference features (No. 7, 8 in Table 2).

Table 4: Impact of features in the LCL predictor

Feature Description	Accuracy
Feature Vector 1 (6 relation status features)	52.8%
Feature Vector 2 (2 sexual preference features)	65.2%
Feature Vector 3 (4 'looking for' features)	54.2%
Feature Vector 4 (all the 12 features)	66.9%

Table 5: Impact of friends number in the FRND predictor

NYC friends # threshold	Accuracy
1	60.0%
5	60.8%
10	61.4%
20	61.8%

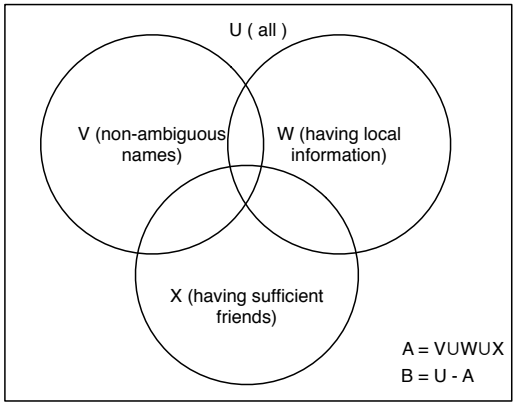

Figure 4: Illustration of different sets

3. *Feature Vector 3* is composed of 4 'looking for' features (No. 9-12 in Table 2).

4. *Feature Vector 4* is composed of all the 12 features (No. 1-12 in Table 2).

Our inference results are shown in Table 4. We can see that *Feature Vector 2* can lead to the highest accuracy among the first 3 feature vectors. This result is perhaps not surprising because sexual preference is generally more correlated to gender than relation status and what people are looking for. This observation will help us improve our following inferences.

6.2.2 Impact of friends number in the FRND predictor

We try to determine the performance of the FRND predictor on users with different number of friends. We generate four training set containing the users who have no less than 1, 5, 10 and 20 NYC friends, and then apply the FRND predictor to them. The results is shown in Table 5. We can see that increasing the NYC friends number threshold from 1 to 20 provides small gains.

6.2.3 Validating the FB predictor Using Boston Network

We validate our FB predictor using another network – "Boston, MA" network (Boston). We first collect the IDs of users who have specified their gender from the Boston network. Then we download those users' profile pages and store their genders and names into a database. We finally obtain 156, 940 users in Boston Facebook, in which there are 53.7% males and 46.3% females.

Since in the Boston database, we only crawled users' names [3], so for each user, we apply the FB predictor to predict the gender. The prediction accuracy is 92.7%. We find that 144, 946 users' names in Boston can be found in our Facebook generated name list. These results show that the FB predictor performs well on and extends to other networks beyond NYC.

7 Inferring Gender for NYC Facebook users

In Section 6, we reviewed the effectiveness of our seven gender predictors on various training sets. Although using a single predictor can achieve a good accuracy, we can further partition the user into various subsets, and apply different predictors to each of these subsets (as illustrated in Figure 4), in order to get better results.

We now present our gender inference method for the entire NYC network. We first provide the approach to partition the users into two groups, A and B. For users in Group B which consists of only 4.9% of NYC users, we have difficulties predicting their gender and have to randomly guess. However, the users belonging to Group A are further divided into various subsets, and we are able to apply our gender predictors to each of these subsets. Finally, we provide our inference results and ideas to further improve the inference accuracy.

7.1 User Partitioning

Inspired by the analysis of our Facebook generated annotated name list presented in Section 4.2, we first partition the users into two groups. For the first group A, we are mostly certain about users' genders, and for the second group B, we are randomly guessing. Users belonging to Group B should satisfy all the following conditions:

- Names never appeared in our annotated name list;
- Do not specify their local information;

- Have very few friends in NYC (we will set a friend number threshold later).

Our detailed partitioning method is described as follows. Let U be the set of all users. Let V be the set of users who have a name in our name list and are not in the ambiguous gender group, i.e., with an $f_m > T_1$ or $f_m < 1 - T_1$, where f_m is the fraction of times that a specific name is labeled as male, and the ambiguous threshold T_1 is in $(0.5, 1]$. Let W be the set of users in U who specified their local information. Let X be the users who have no less than T_2 friends in NYC, where T_2 is a threshold for number of friends. So, we divide the users into two groups: Group A consists of the set $V \cup W \cup X$, and Group B consists of the rest, i.e., $U - A$. Figure 4 presents an illustration of various sets.

7.2 Applying Gender Predictors to Group A

We adopt different gender predictors (discussed in Section 5) to various subsets of users belonging to Group A.

1. For users in $V \cap W$, since their names can be found in the non-ambiguous group of our name list, and have specified their local information, we can adopt the FB-LCL predictor to achieve a high prediction accuracy.

2. For users in $V - W$, whose names can be found in our name list but have not specified local information, by using the FB predictor, we will achieve a high prediction accuracy, if we set an appropriate value for the threshold T_1.

3. For users in $W - V$, it is not effective to use only the name-list based predictors, since their names either have never been labeled or exist in the ambiguous name group. We instead employ a local information based predictor – LCL – for users belonging to the set $W - V$.

4. For users in set $X - V - W$, it is not effective to use the name-list based or local information based predictors. We can, however, predict users' genders using the FRND predictor.

7.3 Gender Inference Results

7.3.1 Parameter Selection

In our experiments, we consider two different thresholds: $T_1 = 0.65$ and $T_1 = 0.8$. We place the users from U, who specified their sexual preference information, in the set W, based on the result in Section 6.2.1. Then, we choose $T_2 = 5$, based on the results from Section 6.2.2. We eventually get a Group A which consists of 96.3% of the users, when $T_1 = 0.65$ and 95.1% of the users when $T_1 = 0.8$.

We then adopt our gender predictors to those users in Group A. We choose inference dataset from users who have identified their genders, and split the dataset into training set and test set by randomly marking each user's gender as unknown with a probability 50%. We list the size of each inference dataset in Table 6.

7.3.2 Results

Table 6 provides a summary of our inference results. In addition to accuracies, we also indicate the fractions of the users belonging to various sets, for the two threshold values $T_1 = 0.65$ and $T_2 = 0.8$. We find that for $T_1 = 0.65$, Group A consists of 96.3% of users and has an accuracy of 95.5%. Also, for $T_1 = 0.8$, Group A consists of 95.1% of users and has an accuracy of 95.2%. These results represent a significant improvement over recent results of gender prediction of [16], which achieved a maximum accuracy of 77.2% based on users' group affiliations. After final inference, the male and female composition of the NYC Facebook network turns out to be 49.8% and 50.2%, respectively. This composition is different from the composition prior to our inference, which is 51.5% males and 48.5% females.

We note that recently Facebook has updated its privacy setting [2]. Under the new default setting, most personal attribues, such as relationship status, "interested in", and "looking for", are only visible to users friends. Though there are fewer information we can get from the new Facebook, our inference method can work well, however. This is because we can still visit users' profile pages, and obtain their names and friend lists. We can then apply our name-list based predictor and friends' information based predictor (such as FB, FRND, and FB-FRND predictors) to predict their genders.

8 Conclusions and Future work

The focus of this paper was on Facebook names and name-centric gender inference.

By crawling Facebook public profile pages for 1.67 million users in New York City, we create a comprehensive and contemporary name list, in which each name is annotated with a popularity estimate and a gender probability. We studied the properties of this annotated name list, and compared it with a popular name list that has been obtained via offline mechanisms. After combining nicknames with their "canonical names," we found that the resulting name popularity has a Zipf distribution, and

Table 6: Accuracies of Gender Inference

Group	Fraction of Users with $T_1 = 0.65$	Training and test dataset size with $T_1 = 0.65$	Accuracy with $T_1 = 0.65$	Fraction of Users with $T_1 = 0.8$	Training and test dataset size with $T_1 = 0.8$	Accuracy with $T_1 = 0.8$
$V \cap W$	21.1%	244,438	97.3%	20.2%	234,562	98.6%
$V - W$	68.1%	365,006	96.8%	65.4%	350,023	98.5%
$W - V$	2.69%	30,195	89.7%	3.54%	40,073	89.6%
$X - V - W$	4.4%	39,712	61.7%	5.94%	54,693	63.0%
A	**96.3%**	**679,351**	**94.6%**	**95.1%**	**679,351**	**95.2%**

that more than 94% of the names can be assigned a specific gender with high probability.

Based on our name list, we developed a new and powerful technique for inferring gender for users who do not explicitly specify their gender. Our name-centric approach to gender prediction partitions the users into two groups, A and B, and is able to accurately predict gender for users belonging to A. Applying our methodology to NYC users in Facebook, we are able to achieve an accuracy of 95.2% for group A consisting of 95.1% of the NYC users. This is a significant improvement over recent results of gender prediction in [16], which achieved a maximum accuracy of 77.2% based on users' group affiliations.

In the future, it will be interesting to study how our method can be extended to datasets from other countries and regions, and how names can be used to predict users' age and ethnicity (both these attributes have an obvious correlation with names). Having inferred the gender of most users in our Facebook dataset, we also plan to learn gender characteristics and study the behavior of males and females in Facebook.

We also plan on exploring how Facebook users' last names, in addition to their first names, can potentially be used to launch a different class of attacks. Using the first and last names, an attacker can extract or brute-force usernames (for popular web/email accounts, e.g., the Facebook account itself) corresponding to their users. Note that typically usernames are derived from first and last names. After enumerating possible usernames extracted from this list, the attacker can possibly launch dictionary attacks on passwords. Second, the usernames can be used to generate possible email addresses that can be exploited for spamming, targeted (e.g., gender specific) advertising and phishing attacks.

References

[1] Facebook statistics. Available at: `http://www.facebook.com/press/info.php?statistics`.

[2] Facebook updates privacy settings. Available at: `http://blog.facebook.com/blog.php?post=197943902130`.

[3] Facebook's gender trouble. Available at: `http://www.salon.com/mwt/broadsheet/2008/07/09/gender_neutral/index.html`.

[4] Popular baby names. Available at: `http://www.ssa.gov/OACT/babynames/`.

[5] CARTER JERNIGAN, B. F. M. Gaydar: Facebook friendships expose sexual orientation. *First Monday 14*, 10 (2009).

[6] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. The weka data mining software: an update. *SIGKDD Explor. Newsl. 11*, 1 (2009), 10–18.

[7] HE, J., CHU, W. W., AND LIU, Z. Inferring privacy information from social networks. In *ISI* (2006), pp. 154–165.

[8] HEATHERLY, R., KANTARCIOGLU, M., THURAISINGHAM, B., AND LINDAMOOD, J. Preventing Private Information Inference Attacks on Social Networks. Tech. Rep. UTDCS-03-09, University of Texas at Dallas, 2009.

[9] KRISHNAMURTHY, B., AND WILLS, C. E. Characterizing privacy in online social networks. In *WOSN* (2008).

[10] KRISHNAMURTHY, B., AND WILLS, C. E. On the Leakage of Personally Identifiable Information Via Online Social Networks. *WOSN* (2009).

[11] LINDAMOOD, J., AND KANTARCIOGLU, M. Inferring Private Information Using Social Network Data. Tech. Rep. UTDCS-21-08, 2008.

[12] MCCALLUM, A., AND NIGAM, K. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization* (1998).

[13] MISLOVE, A., VISWANATH, B., GUMMADI, K. P., AND DRUSCHEL, P. You are who you know: Inferring user profiles in online social networks. In *WSDM* (2010).

[14] QUINLAN, J. R. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research 4* (1996), 77–90.

[15] XU, W., ZHOU, X., AND LI, L. Inferring Privacy Information via Social Relations. In *24th International Conference on Data Engineering Workshop* (2008), pp. 154–165.

[16] ZHELEVA, E., AND GETOOR, L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *WWW* (2009).

Notes

[1] This feature has been removed by Facebook since Aug 2009.

[2] e.g.,`http://www.yeahbaby.com/`,`http://www.moonzstuff.com/articles/nicknames.html`

[3] At the time of the crawl, Facebook has removed the feature that publicly accessing profile pages of users in the same network.