What's in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook

Cong Tang[§], Keith Ross[†], Nitesh Saxena[†], and Ruichuan Chen[‡]

[§]Institute of Software, EECS, Peking University, China
MoE Key Lab of High Confidence Software Technologies (PKU), China
Email: tangcong@infosec.pku.edu.cn
[†]CSE, Polytechnic Institute of NYU, Brooklyn, USA
Email: {ross, nsaxena}@poly.edu
[‡]MPI-SWS, Kaiserslautern, Germany
Email: rchen@mpi-sws.org

Abstract. In this paper, 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.

First, 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 [14], which achieved a maximum accuracy of 77.2% based on users' group affiliations.

Second, having inferred the gender of most users in our Facebook dataset, we learn several interesting gender characteristics and analyze how males and females behave in Facebook. We find, for example, that females and males exhibit contrasting behaviors while hiding their attributes, such as gender, age, and sexual preference, and that females are more conscious about their online privacy on Facebook.

1 Introduction

The 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.

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[14]. 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.

Prior work has considered this problem in the context of Facebook and other OSNs [14]. 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. Similarly, 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 [14]. 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 [14].

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 analyze users' names.

Our Contributions: We make three-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:* 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 [14], which achieved a maximum accuracy of 77.2% based on users' group affiliations.

• *Gender Behavior and Characteristics:* Having inferred the gender of most users in our Facebook dataset, we learn several interesting gender characteristics and analyze how males and females behave in Facebook. We find, for example, that females and males exhibit contrasting behaviors while hiding their attributes, such as gender, age, and sexual preference, and that females are more conscious about their online privacy on Facebook.

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 not aware of any prior research that analyzes and builds on users' names over OSNs (our first contribution).

Zheleva and Getoor [14] proposed techniques to predict the private attributes of users in four real-world datasets (including Facebook) using general relational classi-

fication 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 [14]. 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, 13, 9, 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 [4] 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. [11] 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

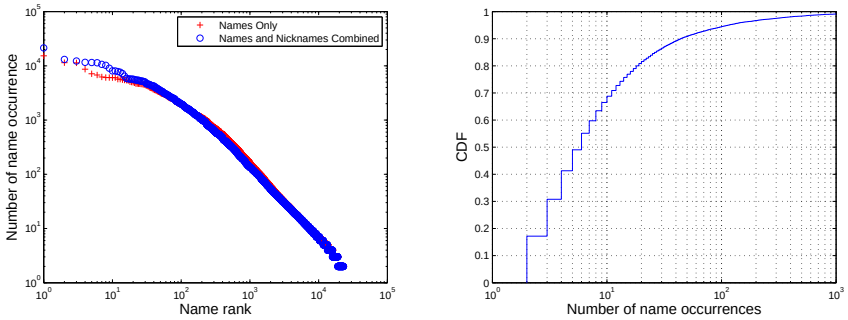
We develop a multi-threaded crawler that visits Facebook user profile pages and stores these pages in a file system. In July, 2009 we obtained a list of Facebook IDs of users in NYC ("New York, NY" network). We were able to do that because at that time, users, by default, were assigned to geographical networks. For each ID, we visit each of its friends, then each of its friends' friends, and so on, until we obtain all NYC users reachable. 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.

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.

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



(a) Distribution of Occurrence of Names (b) CDF of names' occurrence frequency

Fig. 1. Properties of names

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.

We first extract the first names for each of the 1.28 million users and create a crude annotated name list. Note that a Facebook user can choose to "Display Full Name" either as "First Last" or "Last First". We carefully handle this issue. 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 23,363 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 (e.g., `http://www.yeahbaby.com/`, `http://www.moonzstuff.com/articles/nicknames.html`). 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 $x = x + z \cdot \frac{x}{x+y}$, and $y = y + z \cdot \frac{y}{x+y}$. After combining nicknames with names, we obtain a name list with $22,878$ 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 [3] 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 $1,000$ 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 $22,878$ entries, it is much larger and more diverse than the name list we get publicly from [3]. We now study several interesting properties of this name list.

Popularity of names Figure 1(a) 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 1(b), 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.

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 2, it is clear that most names are associated with a specific gender; only about 6% of names

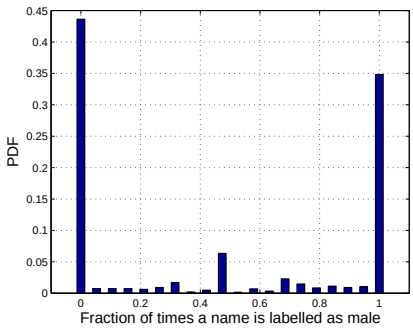


Fig. 2. PDF of fraction of times a name is labeled as male

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. But 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.

5 Design of Gender Predictors

In this section, we propose seven 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).

Besides name list, we take into account additional information, such as users' local and friends information, to improve our prediction. We adopt machine learning algorithms to classify users based on gender. 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 [3], 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.

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 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, which are six relationship status (single, in a relationship, engaged, married, it's complicated, and in a open relationship), two sexual preference settings (interested in men/women), and four "Looking for" attributes (looking for friendship, dating, relationship and networking). Each (binary) feature has a value of 1 if the user corresponds to this feature, and 0 otherwise. For example if the feature "Relationship 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. For the friends who have not specified their gender, we pre-assign genders to them using FB predictor.

5.5 Hybrid Predictors

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

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

Name List, Local and Friend Information Predictor (FB-LCL-FRND) 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) [10] 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 [12].

6.2 Effectiveness of Gender Predictors

We outline our inference results as follows.

- 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.
- LCL predictor provides a higher accuracy (66.9%) than FRND predictor (60.0%) in classification based gender inference.
- 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 (94.6% accuracy), because there is some noise along with the friends' information that decreases the prediction accuracy.

Impact of Features in the LCL predictor We run experiments to determine the local features 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 relationship status features of the user whose gender is to be predicted.
2. *Feature Vector 2* is composed of 2 sexual preference features.
3. *Feature Vector 3* is composed of 4 'looking for' features.
4. *Feature Vector 4* is composed of all the 12 features.

From the results we can see that *Feature Vector 2* can lead to the highest accuracy (66.9%) among the first 3 feature vectors (which are 52.8%, 65.2% and 54.2% sequentially). This result is perhaps not surprising because sexual preference is generally more correlated to gender than relationship status and what people are looking for. This observation will help us improve our following inferences.

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 accuracies are 60.0%, 60.8%, 61.4%, 61.8% sequentially. We can see that increasing the NYC friends number threshold from 1 to 20 provides small gains.

Validating the FB predictor Using Boston Network We validate our FB predictor using another network – "Boston, MA" network (Boston). We 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 (At the time of the crawl, Facebook has removed the feature that publicly accessing profile pages of users in the same network.), 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

We first provide the approach to partition the users into two groups, A and B. The users belonging to Group A are further divided into various subsets, and we are able to apply different gender predictors to each of these subsets to get better results than using a single predictor. For users in Group B we have to randomly guess. 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., $U - A$.

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

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. Then, we choose $T_2 = 5$, based on the results from Section 6.2. 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 $T_1 = 0.8$.

Table 2. 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%**

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 2.

Results Table 2 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 $T_1 = 0.8$. 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 [14], 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 settings [2]. Under the new default settings, most personal attributes, such as relationship status, "interested in", and "looking for", are only visible to users' friends. Though there is now less default information in Facebook, Our inference method continues to work well. This is because we can still visit users' profile pages, and obtain their names and friend lists. Once we obtain the name and friend list, we can predict users' genders.

8 Gender Characteristics and Behavior

8.1 Privacy of Attributes

We apply our inference method to each user in Group A (as discussed in previous section) with parameters $T_1 = 0.8$ and $T_2 = 5$. We compute the percentage of male and female users who hide several of their attributes. The results are shown in Figure 3. Based on the two-proportions z-tests, we confirm that there is a highly significant ($p < 0.0001$) effect of gender on the privacy of each attribute (females showing a higher tendency to hide their attributes). In other words, a larger fraction of females hide their attributes such as gender, age, birthday and relationship status, compared to the male users. Thus, we can conclude that females are more conscious (and intuitively so) in terms of their online privacy on Facebook than their male counterparts. We also examine possible correlations between the hiding of different attributes. For both males and females, we calculate the pairwise Pearson's correlation coefficients, as shown in Table 3. In Social Sciences, correlation coefficients ranging from -0.3 to -0.1 and 0.1

Table 3. Pairwise correlation coefficients for private attributes

Attribute Pair	Males	Females	Attribute Pair	Males	Females
Gender, age	**0.539**	**0.523**	Age, looking for	0.244	0.252
Gender, birthday	**0.731**	**0.77**	Birthday, relationship status	0.486	**0.51**
Gender, relationship status	**0.51**	0.5	Birthday, sexual preference	0.407	0.383
Gender, sexual preference	0.433	0.376	Birthday, looking for	0.392	0.432
Gender, looking for	0.437	0.444	Relationship, sexual preference	**0.582**	**0.561**
Age, birthday	**0.738**	**0.669**	Relationship status, looking for	**0.625**	**0.681**
Age, relationship status	0.325	0.311	Sexual preference, looking for	**0.579**	**0.558**
Age, sexual preference	0.284	0.265			

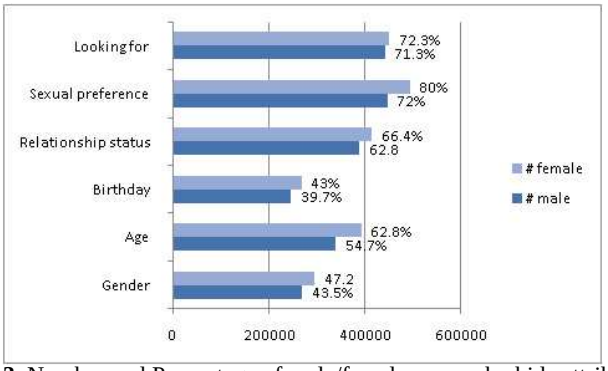


Fig. 3. Number and Percentage of male/female users who hide attributes

to 0.3 are generally regarded as small, -0.5 to -0.3 and 0.3 to 0.5 as medium, and coefficients larger than 0.5 and smaller than -0.5 as high [5]. From the results in Table 3, we find that the strongest correlations are between "hiding of gender" and "hiding of birthday", and "hiding age" and "hiding birthday", both for males and females. This is followed by correlations between "hiding relationship status" and "hiding looking for", and "hiding relationship status" and "hiding sexual preference". "Hiding of relationship status" and "hiding of sexual preference", and "hiding of gender" and "hiding of age" are also strongly correlated. These correlations are more or less consistent for both males and females, and imply that users who hide one attribute is also likely to hide several of other attributes. Looking further into these correlations, we find two independent clusters consisting of private attributes for both genders: (gender, age, birthday) and (relationship status, sexual preference, looking for).

8.2 Targeted Advertising and Privacy Implications

We provide examples of how third parties could use our results for gender-specific online stalking and targeted advertising. These third parties can use the resulting gender information from our gender inference methods combined with users' attributes, to help improve the accuracy of targeted advertising.

For example, an online dating company might be very interested in marketing their services and websites to single males and females who are looking for dating. In NYC Facebook, we find that there are $35,076$ males and $14,865$ females matching this criteria. A cosmetic company might be interested in marketing there products to young females, while we find that there are $106,007$ females that are in their 20s in NYC Facebook. A "gifts for lovers" company might want to know the information of people that are in a relationship. Our statistics show that there are $46,522$ males and $48,328$ females who specified that they are in a relationship.

There are several other interesting and concerning implications of our results. For example, there are 752 males and 463 females in NYC Facebook who are married but looking for dating; there are $9,077$ males indicating their sexual preference as men, however, 18.2% of them are in a relationship, and 5.88% of them are married; similarly, there are $18,259$ females specifying their sexual preference as women, but 18.3% of them are in a relationship, and 17.1% of them are married. All these statistics and others can potentially be used by malicious parties with unknown nefarious intentions.

9 Conclusions

The focus of this paper was on Facebook names, name-centric gender inference and gender behavior. By crawling Facebook public profile pages for 1.67 million users in New York City, we create a comprehensive and contemporary name list. We studied the properties of this annotated name list, and compared it with a popular name list that has been obtained via offline mechanisms. Based on our name list, we developed a new and powerful technique for inferring gender for users who do not explicitly specify their gender. 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. Having inferred the gender of most users in our Facebook dataset, we learn gender characteristics and analyze how males and females behave in Facebook.

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. Popular baby names, available at: `http://www.ssa.gov/OACT/babynames/`
4. Carter Jernigan, B.F.M.: Gaydar: Facebook friendships expose sexual orientation. First Monday 14(10) (2009)
5. Cohen, J., Cohen, P., West, S., Aiken, L.: Applied multiple regression/correlation analysis for the behavioral sciences. Erlbaum Hillsdale, NJ (1983)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. (2009)
7. He, J., Chu, W.W., Liu, Z.: Inferring privacy information from social networks. In: ISI. pp. 154–165 (2006)
8. Heatherly, R., Kantarcioglu, M., Thuraisingham, B., Lindamood, J.: Preventing Private Information Inference Attacks on Social Networks. Tech. Rep. UTDCS-03-09, University of Texas at Dallas (2009)
9. Lindamood, J., Kantarcioglu, M.: Inferring Private Information Using Social Network Data. Tech. Rep. UTDCS-21-08 (2008)
10. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: AAAI-98 Workshop on Learning for Text Categorization (1998)
11. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: Inferring user profiles in online social networks. In: WSDM (2010)
12. Quinlan, J.R.: Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research 4, 77–90 (1996)
13. Xu, W., Zhou, X., Li, L.: Inferring Privacy Information via Social Relations. In: 24th ICDE Workshop. pp. 154–165 (2008)
14. Zheleva, E., 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)