Attacking Age Privacy in Online Social Networks

Ratan Dey[1], Cong Tang[2], Keith Ross[1], and Nitesh Saxena[1]

[1] Polytechnic Institute of New York University, Brooklyn, US
[2] Peking University, Beijing, China

Abstract. Birth year is a fundamental human attribute, and for many people a private one. We have found that in our sample dataset of 1.47 million Facebook users from New York City, only 1.5% of them specify their age in their public profile, confirming that age is indeed a private attribute for most users. In this paper, we investigate whether it is possible to estimate the age of each of the remaining 98.5% of the the New York City Facebook users. To estimate Facebook user ages, we develop a novel two-step procedure. In the first step, we exploit side information such as high-school graduation year and high-school graduation years of friends with the same high school name to accurately estimate the age for a large set of users. In the second step, we exploit the underlying social network structure to design an iterative algorithm, which derives age estimates based on friends's ages, friends of friends' ages, and so on. Our overall methodology is able to estimate age of 84% users with a 4-year mean absolute error. However, we find that for many older users, age is difficult to estimate accurately, and may thus remain private within OSNs. We also develop a technique for another related privacy violation – classifying a user as a minor (under 18 years of age) or as an adult. Our work casts serious doubts on age privacy and children online privacy in OSNs.

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 – let's call her 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 and profile picture. In particular, Alice has the option of hiding her attributes, such as age, 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 only 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, relationship status, and political affiliation that Alice has not explicitly made public [18]. To the extent this is possible, third parties not only can use the resulting information for online stalking and targeted advertising, but could also sell it to others with unknown nefarious intentions.

In this paper, we take an in-depth look at the age (birth year) privacy of Facebook users. We have found that in our sample dataset of 1.47 million Facebook users from

New York City, only 1.5% of them specify their age in their public profile, confirming that age is indeed a private attribute for most users. Motivated by this, we ask the question: *whether it is possible to estimate the age of the remaining users – i.e., those who aim to hide their ages – with a high accuracy?* We seek to answer this question using algorithms that are not Facebook specific, so that they can be applied to OSNs in general. For age estimation, we only use public profile and friendship information; we do not use image analysis or network/group information.

As a related subject, we also consider the problem of identification of minors among Facebook users. We define a minor to be a person who has reported to Facebook his or her age as less than 18 years. We note that Facebook is very popular among minors. For example, there are many high school networks in Facebook to which most of the minors belong. Facebook provides a virtual world where minors, similar to other users, can keep in touch with their friends and share among friends and have many other opportunities. However, at the same time, minors can be victims of different types of crimes such as cyber-bullying or sexual harassment. As reported in [7], there is ,"More cyber-bullying on Facebook, and social sites than the rest of of Web". There has also been an incident of cyber-bullies harassing a teen on Facebook before , after her suicide [1]. Most OSNs, including Facebook, therefore incorporate mechanisms in an attempt to protect minors from online criminals and to enforce their online privacy. As an example, minors do not show up on public search results. We pose therefore the following question: *Is it possible to classify Facebook users as adults and minors?*

In this paper, our technical contributions are as follows:

- *Large Data Set:* We crawled Facebook to get two large Facebook data sets, both of which targeting the New York City (NYC) Facebook users. In July 2009, we crawled all 1.69 million users in NYC, obtaining Facebook IDs and their full profile pages. Many of users in this July 2009 dataset explicitly provide their age, thereby allowing us to create an extensive ground-truth test set. In March 2010, we launched another extended crawl, during which we visited the 1.67 million NYC user IDs from the July 2009 dataset. Among the 1.67 million user IDs, we were able to revisit 1.47 million of the users. At the time of March 2010 crawl (in fact since early 2010), due to changes in Facebook's default privacy settings, one can only crawl *limited profile pages* of Facebook users, even if one lives in the same geographical region. Only this limited profile information is available to an attacker today. We found that only 82.73% of the limited profile pages publicize friend lists, and a mere 1.5% of them provide the users' ages.
- *Age estimation:* Our primary goal is to estimate the age of all NYC Facebook users, based only only the limited profile information provided in March 2010 dataset. We develop a novel two-step estimation methodology. In the first step, we exploit side information such as high-school graduation year and high-school graduation years of friends with the same high school name to accurately estimate the age for a large set of users. In the second step, we exploit the underlying social network structure to derive an iterative algorithm, which derives age estimates based on friends' ages, friends of friends' ages, and so on. We also propose using *reverse friend lookup* to determine the friends of a user when that user hides his friend list. Our overall methodology is able to estimate age of 84% users with a 4-year mean absolute error.

However, we find that for many older users, age is difficult to estimate based only public profile information and friendship links only. We examine why it is more difficult to estimate the age of older users. To the best of our knowledge, this paper is the first in-depth study of the age estimation problem in OSNs.

– *Minor classification:* Facebook takes special precautions to protect the privacy of a minor, including not allowing a minor to explicitly indicate in his public profile page his age or even that he is a minor. We investigate whether an attacker can determine this private information. Specifically we develop a technique for classifying a user as a minor or as an adult. This paper represents, to the best of our knowledge, the first systematic study for minor identification in an OSN. Using some Facebook-specific features, our analysis shows that one can classify a large majority of users, with a high-degree of certainty, as either adult or minor.

We note that our age inference approach is simple enough for naive attackers to develop and execute with effect. This implies that Facebook age privacy can be violated very easily. We believe that our work casts serious doubts on age privacy and children online privacy in OSNs.

The remainder of this paper is organized as follows. We present our data gathering mechanism and properties of the dataset in Section 2. Next, we present our age estimation methodologies and results in Section 3 and Section 4. In Section 5, we describe the step by step methodologies for classifying NYC dataset users as minors or adults. In Section 6, we discuss relevant prior work. Finally, Section 7 summarizes our conclusions.

2 Data Sets and Performance Measures

2.1 Crawling NYC Users and Their Friends

In Facebook, when Alice visits Bob's profile page, the information that is displayed to her depends on her relationship with Bob (for example, whether she is a friend or not) and on Bob's privacy settings. Roughly speaking, when Alice is a Facebook friend of Bob, then she typically gets to see Bob's **full profile page**, which includes all of Bob's friends as well as all of the information and photos that Bob puts into Facebook; if Bob is not a friend, Alice only gets to see a **limited profile page**, which often includes no more than Bob's full name and his photo.

We developed a multi-threaded crawler that visits Facebook user profile pages and stores the pages in a MySQL database. Using this crawler, in July 2009, we crawled all the users in NYC, obtaining their Facebook IDs and their *full profile pages*. We were able to do this because at that time (i) users were, by default, assigned to regional networks; and (ii) a user's full profile page was, by Facebook's default privacy setting, made public to all other users in the same network. We obtained 1.67 million NYC user IDs and their corresponding full profiles. We refer to this dataset as the *July 2009 dataset*. Facebook fully deprecated regional networks as of late September 2009 [5, 2]. A user's full profile is now, by default, only available to the user's friends.

In March 2010, we launched another extended crawl, during which we visited the 1.67 million NYC user IDs from the July 2009 dataset. Among the 1.67 million user IDs, we were able to re-visit 1.47 million of the users; the remaining accounts appear to have been deactivated or removed by Facebook between our two crawls. At the time of

March 2010 crawl, due to changes in Facebook's default privacy settings, we obtained the *limited profile pages* of the NYC users. As shown in Table 1, only 82.73% of the limited profile pages publicize friend lists, and a mere 1.5% of them provide the users' ages.

During the March 2010 crawl, for each crawled user (say, Alice), in addition to obtaining Alice's limited profile page, we also collected the limited profile pages of her friends, whenever she made the friend list publicly available. By crawling the friends of the 1.47 million NYC users, we obtained an additional 47.79 million users, many of whom do not reside in NYC. Our *March 2010 dataset* has the limited profile pages of 49.26 million users, consisting of the 1.47 million NYC users and their friends. This data set contains approximately 306 million friendship links between NYC users and their friends. We emphasize that the data set does not include the friendship links between non-NYC users, as that would have required significantly more computational and bandwidth resources than available at the time. *Our primary goal is to estimate the ages of the* 1.47 *million NYC users using the data (limited profiles) in the March 2010 data set.* The July 2009 dataset, containing full profile pages, is used for ground truth and evaluation of the methodology.

Table 1. Properties of the March 2010 Dataset (containing limited profiles)

Property name	**Value**
# users in NYC	$1,473,199$
# users' friends outside NYC	$48,828,008$
% users who do not make friends public	17.27%
% users who specified age	1.5%
% users who make HS graduation year public	21.6%
% users who provide work place network public	3.7%
% users who provide grad/college info public	19.0%

2.2 Reverse Friend Lookup

As shown in Table 1, a significant fraction of users do not disclose their friend lists in their limited profiles. It is, however, possible to obtain partial friend lists for such users employing a reverse lookup mechanism. Specifically, if Bob hides his friend list, we can look at all other users who disclose their friend lists, and identify those who indicate they are friends with Bob. We remark that such a friend list for Bob is incomplete, as it only contains friends who both (i) reside in NYC, and (ii) do not hide their friend lists. Figure 1 shows the fraction of users among those hiding their friend lists for which reverse lookup can identify x friends. For example, for 46.26% of these users we can find at least 15 (NYC) friends. Clearly, with a more extensive crawl, which would also obtain the friend lists of the non-NYC users, reverse lookup would yield a much more complete view of these otherwise hidden friend lists.

2.3 Inactive Users

Although many Facebook users have hundreds of friends and 50% of users visit the site daily (as discussed in [4]), many accounts have few friends and no recent activity; we refer to such dormant users as *inactive users*. In order to prevent these users from skewing the results of our study, we do not attempt to estimate the ages of users who

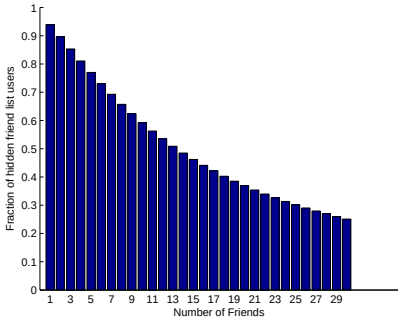


Fig. 1. Fraction of users for whom reverse lookup can identify x friends

satisfy all of the following conditions: (i) the user has 10 or fewer friends; (ii) the user does not provide his or her birth year. (iii) the user does not provide high-school graduation year. That is, we do attempt to try to estimate the age of low activity users, unless they explicitly provide their age or their high-school graduation date. After removing all users who do not satisfy any of the above three criteria, we have $1,191,758$ NYC users, for whom we will attempt to estimate their ages.

2.4 Estimation Performance Measures

In order to evaluate the performance of our age estimation procedures, we utilize two different measures: the Mean Absolute Error (MAE) and the Cumulative Score (CS). MAE is defined as the average of the absolute differences/errors between the estimated ages and "ground truth" ages, i.e., MAE $=\sum_{k=1}^{N} |x'_k - x_k|/N$, where x_k is the ground truth age for the user k, x'_k is the estimated age, and N is the total number of test users. The MAE measure has previously been used in the context of age estimation based on facial images [10–12] (reviewed in Section 6). The cumulative score, on the other hand, is defined as $CS(j) = N_{e \leq j}/N \times 100\%$, where $N_{e \leq j}$ is the number of test users for which the age estimation procedure makes an absolute error no higher than j years. For example, $CS(4)$ is the percentage of test users for which the absolute error is less than 4 years. This measure has previously been used in [10].

For calculating MAE and CS, we use the birth year data from the July 2009 dataset as ground truth. As described earlier, while crawling Facebook in July 2009, by default, we were able to obtain the full profile pages of the users in NYC. In the July 2009 data set, $515,000$ users provide their birth years. In the second crawl (March 2010), we found that $486,686$ of these user accounts were still active. However, some users blatantly lie about their ages, reporting ages over 80 when they are actually much younger. We therefore remove from our ground-truth data set any user who reports a birth year prior to 1931 (This step removes a small number of users who are actually over 80) and who is identified as inactive user as discussed in section 2.3. At this stage, we have $419,395$ users' birth years which will be used as ground truth to determine the accuracies of the age estimation methods.

We briefly remark that users can easily lie about their ages in Facebook. However, given that a Facebook user typically has family, high-school and university friends who know with certainty the user's age, it is difficult for an adult user to lie about his age.

Some minors, however, say they are over 18 to get adult privileges. Lying appears to be very difficult to account for in age estimation in OSNs.

3 Birth Year Estimation: Basic Methods

In this and the following section, we present our age estimation methodology. The methodology is based on fundamental attributes of OSNs, i.e., limited profile information and social links, and does not use features that are highly application (Facebook) specific. Let $\mathcal{G}$ be the set of all $1,191,758$ NYC Facebook users for which we will attempt to estimate the birth year. Our approach is to first find a subset $\mathcal{G}_0$ for which we can estimate the birth year with a high accuracy. Then, we find another disjoint subset $\mathcal{G}_1$ for which we can estimate the birth year with somewhat lesser accuracy. Iterating in this manner, we create a partition $\{\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_N\}$ of $\mathcal{G}$ with a different estimation procedure and estimation confidence for each disjoint subset.

3.1 Low Hanging Fruit

The set $\mathcal{G}_0$ is a set of users who make their birth years publicly available in their limited profiles. For a user in this set, we simply estimate the user's birth year as the publicly specified birth year. Assuming that the reported ages are correct, our birth year estimates for the users in $\mathcal{G}_0$ is obviously 100% accurate. The set $\mathcal{G}_0$ consists of $15,975$ users or 1.34%. We denote this trivial age estimation procedure as Step 0.

3.2 Benchmark

We briefly mention here that we experimented with estimating users' birth year using mean and median statistics, such as the mean and median birth year, for the NYC users outside of $\mathcal{G}_0$. The median and mean birth years are 1983 and 1980, respectively; the corresponding MAEs are 8.91, 8.52, respectively. CS versus error level (in years) is depicted in Figure 2(a). From the graph, we can observe that the estimation accuracies are relatively high. Specifically, mean and median statistics can achieve an error within 4 years for only 40% of the users, and an error within 10 years for only 70% of the users. This naive approach provides us with a benchmark to compare the performance of our estimation algorithms.

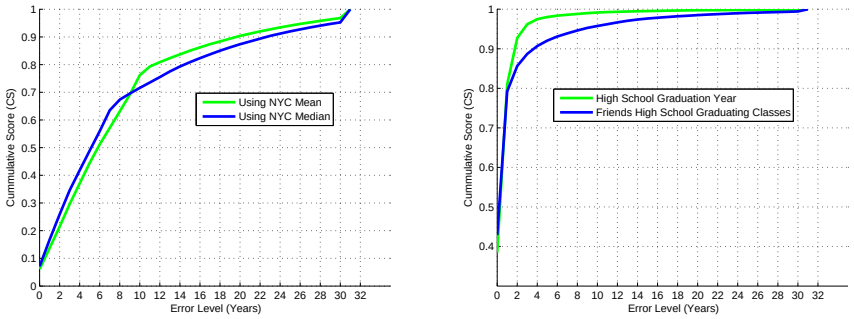
3.3 Using High School Graduation Year

There are many users who do not make their birth year publicly available in their limited profiles, but nevertheless make their high school graduation year publicly available. Because most people complete high school between the ages of 17 and 19 years, the high school graduation year is clearly correlated with the birth year of an individual. To take advantage of this correlation, we build a training set for identifying the relationship between high school graduation year and birth year.

From our March 2010 dataset (including NYC users and their non-NYC friends), we found that $255,012$ users made both their birth year (BY) and high school graduation year (HSY) public. We fed these $255,012$ data points into linear regression and obtained the following regression line:

$$BY = 0.9368 \times HSY + 108.2107 \tag{1}$$

The set $\mathcal{G}_1$ is the set of NYC users who do not make their birth year publicly available, but make their high-school graduation year publicly available. In $\mathcal{G}_1$, there are $215,846$ users representing 18.11% of users in $\mathcal{G}$. Using the Equation 1, we assign birth



(a) Estimating birth year using mean and median

(b) Estimating birth year using high school graduation year and using friends'high school graduating class

Fig. 2. Accuracy of Benchmark and Basic Methods

years for the these $215,846$ users. We refer to this as Step 1. Of these users $215,846$, $98,653$ belong to our ground truth data set, yielding an MAE of 1.11. Figure 2(b) depicts the cumulative score for this linear regression estimation procedure; note that for 94% of the users, the linear regression results in an error of 2 years or less. We remark that many users also provide college and university graduation dates. However, we found college and university graduation dates to be much less reliable estimators of age than high school graduation dates. For that reason, we do not use college and university graduation dates in our estimators.

3.4 Using Friends' High School Graduating Classes

A user may not publicize her birth year or her high school graduation year, but she may have many friends from her high school graduating class from which we may be able to infer her high school graduating year. To create the subset $\mathcal{G}_2$, we use a grouping methodology that takes into account the high school name and graduation year of a user's friends. The methodology is as follows. For each user u not in $\mathcal{G}_0 \cup \mathcal{G}_1$, among u's friends we find the most frequently occurring high school graduating class (i.e., high school name and graduation year). If u has T or more friends in this high school graduating class, we put u in $\mathcal{G}_2$ and assume that u is also from this high school graduating class. Let y_u be the corresponding graduation year. To estimate user u's age, we then use y_u as HSY in the regression Equation 1. We call this procedure Step 2. There are $919,680$ users in $\mathcal{G} - (\mathcal{G}_0 \cup \mathcal{G}_1)$. Using $T = 6$, we find $453,596$ users in $\mathcal{G}_2$. Using $T = 6$ gives us moderate coverage and accuracy (low MAE); if we choose a smaller value for T, coverage improves but accuracy degrades. Of these $453,596$ users, $141,216$ are found in the ground truth verification set. For these $141,216$, the MAE for our estimation procedure is 1.86. Figure 2(b) shows the corresponding cumulative score. We define $\mathcal{H} = (\mathcal{G}_0 \cup \mathcal{G}_1 \cup \mathcal{G}_2)$.

Table 2 summarizes the results from Steps 0, 1, and 2. From these three steps, we have been able to estimate the ages of 57.51% of the users with a high-level of accuracy

Table 2. Summary of results from Steps 0,1,2

Set	Number of users	Percentage	MAE	CS(4)
$\mathcal{G}_0$	$15,975$	1.34%	0	100%
$\mathcal{G}_1$	$215,846$	18.11%	1.11	96%
$\mathcal{G}_2$	$453,596$	38.06%	1.86	91%
$\mathcal{H}$	$685,417$	57.51%	1.5	95%

of MAE 1.5. However, there still remains $506,341$ $(42.49$ %) users outside of $\mathcal{H}$ for which we need to use more advanced procedures to estimate ages.

4 Iterative Method

The method in Section (3.4) makes use of the age distributions of a user's friends; however, it does not take advantage of the underlying network structure in the social network, which provides information about friends of friends, friends of friends' friends, and so on. To exploit this underlying network structure, we develop an iterative algorithm. This iterative algorithm is not limited to age estimation – it can be used to estimate other attributes in social networks as well.

In our algorithm, at each iteration i, we have age estimates for a set of users, denoted $E(i)$. For each user $u \in E(i)$, let $x_u(i)$ be our estimate of u's age at the i-th iteration. Also let F_u be the set of u's friends, and $F_u(i)$ be the set of u's friends for which we have age estimates, that is, $F_u(i) = F_u \cap E(i)$.

In the iteration scheme, for any user $u \in \mathcal{H}$, we set $x_u(i) = a_u$, where a_u is the age determined in the previous section. For a user $u \notin \mathcal{H}$ which has at least one friend with an age estimate (i.e., $F_u(i) \neq \phi$) we use iterations:

$$x_u(i+1) = \alpha x_u(i) + (1-\alpha)\Phi[x_v(i)),\ v \in F_u(i)], \tag{2}$$

where $\Phi[\cdot]$ could be as simple algebraic expression or a more sophisticated clustering algorithm. We will soon provide some examples for $\Phi[\cdot]$. To initialize the iterations, we set $E(0) = \mathcal{H}$. We also set $E(i+1) = E(i) \cup \{u : F_u(i) \neq \phi\}$. Notice that this algorithm takes into account Bob's friends of friends when estimating his age.

Since the function $\Phi[\cdot]$ must be calculated for millions of users at each iteration, it is critical to choose a function that not only provides good estimates but is also computationally efficient. We examine two computationally-efficient approaches in this paper: linear regression and percentiles.

For the linear regression approach, we choose a linear function of the mean, median, and standard deviation of the user's friends; specifically, a function of the form

$$\Phi[x_v(i)),\ v \in F_u(i)] = a_1 MEAN_u(i) + a_2 MEDIAN_u(i) + a_3 STD_u(i) + a_4$$

where $MEAN_u(i)$ (respectively, $MEDIAN_u(i)$ and $STD_u(i)$) is the mean (respectively, the median and standard deviation) of the values in $F_u(i)$. This linear equation is efficient to calculate, but how should we choose the values for a_1, a_2, a_3, and a_4?

We use linear regression to determine the coefficients a_1, a_2, a_3, and a_4. Specifically, for each of the $685,417$ users in $\mathcal{H}$, we determine the mean, medium, and standard deviation of the user's friends' ages. For each user in $\mathcal{H}$, we have a data point consisting of the user's age as well as the associated mean, median and standard deviation. We

feed these $685,417$ data points into a standard linear regression procedure to obtain the values of a_1, a_2,a_3, and a_4. The resulting regression equation becomes:

$$BY = 0.3583 \times MEAN + 0.6654 \times MEDIAN - 0.3596 \times STD - 45.5534 \quad (3)$$

For the percentile approach, with a given value of q, $\Phi[\cdot]$ is simply the the q percentile of the ages in $F_u(i)$. For example, with $q = 70$, we take the age x so that 70% of the users in $F_u(i)$ are younger than x. Note that $q = 50$ is simply the median of the ages in $F_u(i)$. We experimented with using different percentiles such as 50th (median), 60th, 70th, 80th etc and found that 70th percentile provided the best estimation accuracy in terms of MAE and CS.

4.1 Results for Iteration

We first applied the regression equation 3 for the function $\Phi[\cdot]$. If a user has more than 20 friends with known ages, we assign less weight (α) to the new estimates; and if the user has at most 20 friends with known ages, we assign more weight to new estimates, with the hope that some of user's friends will be assigned ages in subsequent iterations. We have set the value $\alpha = 0.6$ for users who have at most 20 friends (with known ages) and $\alpha = 0.90$ for users who have more than 20 friends (with known ages).

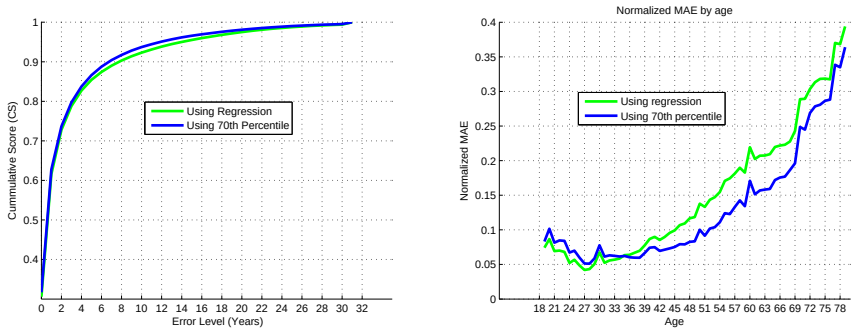
We then applied the 70th percentile of friends ages for the function $\Phi[\cdot]$. Again we modify the value of α depending on how many friends a user has with known ages. We set the $\alpha = 0.6$ for users who have at most 20 friends (with known ages) and $\alpha = 0.90$ for users who have more than 20 friends (with known ages).

There are $506,341$ users in the set $\mathcal{G} - \mathcal{H}$. After running the iterative method for **5** iterations, we were able to assign ages to 505 thousands additional users in both approaches. Of these 505 thousands users, $171,187$ belong to our ground truth data set. Over the set $\mathcal{G} - \mathcal{H}$, iterations with regression gave an MAE of 5.13 and CS(4) of 66.8%, whereas iterations with percentiles gave MAE of 4.48 and CS(4) of 69.3%.

For the remaining few thousand users, we simply use mean birth year (i.e., 1980), which we found to yield better results than the median. Figure 3(a) shows the accuracy of overall methodology (combining basic profile information, reverse friend lookup, and iterations with regression and percentiles). The overall method using iterations with 70th percentile, we obtain an MAE of 2.71 and CS(4) of 83.8%. *Thus, the overall methodology is quite accurate, and is significantly more accurate than the baseline approach of using means or medians for the users outside of* $\mathcal{G}_0$. Our age inference approach is simple enough for naive attackers to develop and execute with effect. This implies that Facebook age privacy can be violated rather easily for most Facebook users.

4.2 Defenses for the Age Privacy Attack

Due to page constraints, we only briefly discuss what a Facebook user, and Facebook, can do to avoid age privacy attacks. The user can configure her privacy settings so that age, high-school graduation year, and friend lists are not available in her limited profile (that is, to non-friends). However, this alone will not fully protect the user, since an attacker can still perform reverse-friend lookup. With reverse friend lookup, the attacker may find a group of friends all from the same high-school graduation class,



(a) Overall accuracy after combining basic and iterative methods

(b) Normalized MAE by age for all users after combining basic and iterative methods

Fig. 3. Accuracy and normalized MAE after combining basic and iterative methods

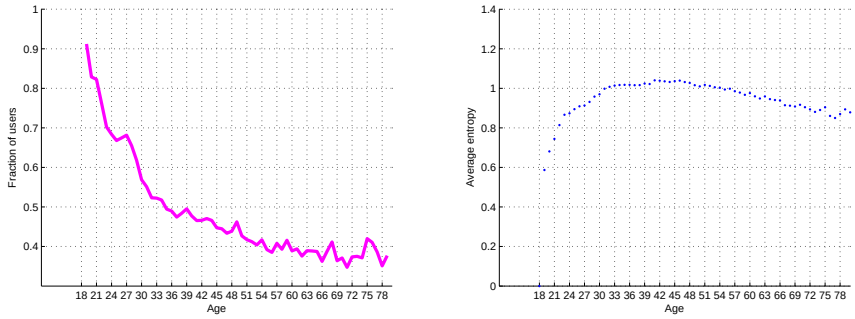
which – as we saw – can provide highly accurate estimates of age. The attacker can also apply iterations, as previously described, to obtain good estimates for age. Note that reverse lookup can also be potentially used to infer not only age, but other attributes including religious and political preferences. To prevent reverse friend lookup, when Alice chooses to hide her friends in her limited profile, Facebook could also automatically remove Alice from the friend lists in all her friends' limited profiles. We strongly recommend that Facebook adopt this policy.

4.3 Age Analysis

We now examine the performance of the iterative method as a function of the age of the users. For each age we determine the **normalized MAE**, which is defined to be as the MAE for all users of that age divided by that age. So, for example, the normalized MAE for 27 is the average MAE for all ground truth users of age 27 divided by 27. Figure 3(b) presents the normalized MAE per age resulting from our methodology (combining the basic methods with reverse friend lookup and iterations with percentiles). We observe that (i) our method has a normalized MAE of under 0.1 for all ages under 50; (ii) after age 50, the performance of our method begins to decline – for example, for users older than 70 the normalized MAE exceeds 0.25.

We now investigate why it is difficult to accurately estimate age for users over 50 when using profile and friendship information. (It may be possible to improve the estimation accuracy by taking additional information into account, such as the users' photos and the networks to which the users belong. Such a study is beyond the scope of this paper.) Figure 4(a) shows, for each age, the fraction of users who provide strong hints about their age (either by explicitly stating their age, or providing their high-school graduation year in their limited profiles). We see that for users under 25, more than 70% in each group provide strong hints. However, for users over 50, less than 40% provide strong hints. Thus, one reason why it is easier to estimate the ages of younger users is that they are more forthcoming about their age (either directly or indirectly through high-school graduation year) in their limited profiles.

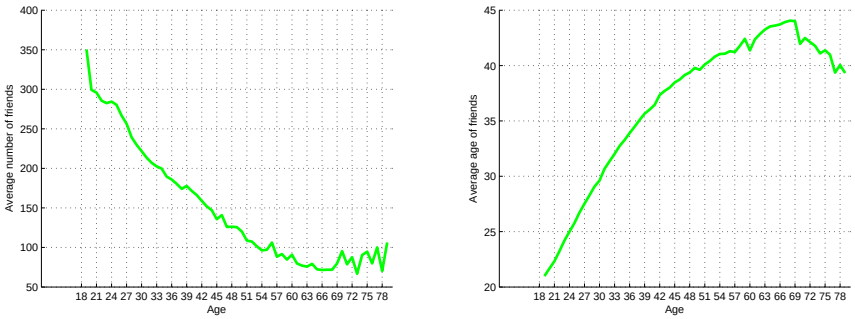
Given that it is hard to estimate the age of an older user directly from the information in his/her limited profile, we then examine how much information is available from



(a) Fraction of users provide birth year or high school graduation year at each age

(b) Average entropy at each age

Fig. 4. Providing age specific information and friend entropy at each age



(a) Average number of friends at each age

(b) Average age of friends' ages (which can be determined from basic methodology) at each age

Fig. 5. Average number of friends and average age of friends' ages at each age

these users' friends. Next we examine the diversity of friends for older users. For each user, we determine the distribution of its friends ages, then the corresponding entropy of the distribution. In Figure 4(b), we have plotted the average entropy at each age on y-axis and age on x-axis. From this plot, we can observe that very young users (ages $18-22$) have low values of entropy, whereas all other users have relatively high values. This greater diversity in friends' ages for older users makes it more difficulty to infer age from the ages of friends.

Figure 5(a) shows the average number of friends for each age group. Here we see a dramatic difference between the younger and older users. In particular, we see that users of age 30 have, on average, more than twice as many friends as users over 50. The fewer the friends a user has, the less the information that is available for a friend-based inference procedure. Figure 5(b) shows the average age of friends' ages (determined from the basic methodology) for each age group. The results here a particularly striking: up until age 50 the curve is almost linear, but after 60 the figure is no longer monotonically

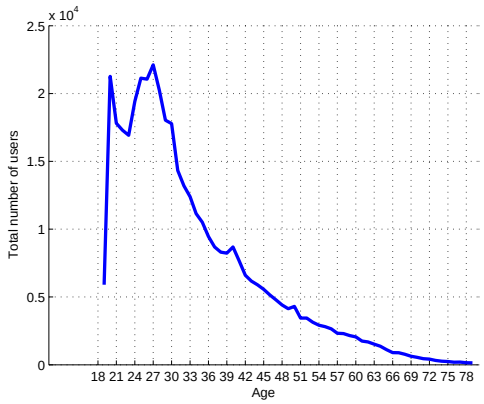


Fig. 6. Total number of users at each age

increasing. Therefore, users over 70 cannot be distinguished from users over 60 based on their friends' ages.

Finally, 6 shows the total number of users for each age group in our ground truth dataset. We see that there are many more younger users than older users. This implies that when a user's age cannot be easily estimated, because there are many similar training profiles for widely different ages, a machine learning algorithm will tend to assign a young age to the user, since there are so many more young users.

In summary, because an older user often does not have many friends, the age diversity of his friends is high, the age distribution of his friends' ages is often similar to those of middle-age users, and the fact that there are many more younger users, it is very difficult to get accurate age estimates for older users based on friendship information. Combining this observation with the fact that older users generally do not give strong hints about their age in their limited profiles, makes the problem of identifying older users in OSNs a very challenging problem.

5 Identifying Minors

We define a person as a minor if he/she is under 18 years old; anyone who is 18 or older is an adult. Facebook minors have a slightly different experience with privacy than adults [6]. When a minor sets information such as age, education information, photos or status updates to be visible to "Everyone," that information is actually only visible to her friends, friends of friends, and people in any school or work networks she has joined. For a minor, only a limited amount of information is available to a non-friend, namely, photo, name, networks joined, and gender.

In this section, we consider the following classification problem: Given a Facebook user whose age is not publicly available, classify the user as a minor or an adult. Unlike for the age estimation problem, for identifying minors we will need to exploit specific features of Facebook as of February 2011.

5.1 Public Search

Facebook allows for public search of a user, that is, visiting the user's public profile without logging on to Facebook [3]. This feature is often used by search engine crawlers (e.g., Google) so that a Facebook user can be searched directly from search engines.

Users can use their privacy settings to opt out from public search. If public search status is enabled for a user, depending on the user's privacy settings, Facebook provides the user's name, profile picture, gender, and a small subset of friends. If public search is disabled, in March 2010, Facebook sent either "Page Not found" (PNF) or redirects the user to a "Sign Up Facebook" (SUF) page. Public search is always disabled for minors (anyone under 18). During our March 2010 crawl, we also collected the status of public search for every user in our data set. Our March 2010 dataset is therefore augmented with the public search status of each of the 49.26 million users.

5.2 Ground Truth

In order to evaluate a minor identification algorithm, we need to obtain ground truth minor and adult users. To this end, we make the following observations, if public search is enabled for a user, then we know for sure the user is an adult. If public search is disabled, then the user may or may not be a minor. In March 2010, for a minor, if the SUF page was returned, then the user is a minor for sure. If the PNF page is returned, depending on the privacy settings of the user, the user may be either a minor or an adult. Using these observations, we were to obtain $19,488$ ground-truth minors and $1,081,567$ ground-truth adults.

Recently, Facebook has changed public search results. At present (February 2011), both for adult users who opt out of public search and for all minors, Facebook returns the PNF error page. *Therefore, the challenge is to determine, for all NYC users for which public search is disabled, which are minors and which are adults. There are* $391,632$ *such users.*

5.3 Step 1: Add him/her as a friend

When we browse for a user in Facebook today, depending on his privacy settings and age, we may see a "Send him/her a message" link in his profile page . In the case of an adult, this link will appear by default (but an adult can have it removed through his privacy settings). In the case of a minor, "Send him/her a message" will never appear in the user's public available profile page. So if we find the "Send him/her a message" link in a public profile page, we say the user is an adult. Using "Send him/her a message" information, we can designate $280,871$ users as adults.

So in this step we have identified adults only, that is, among these 1.47 million users we have identified $1,363,438$ users as adults in step 1. When we collected public search status for each of the user of NYC dataset, we were unable to get any result for $21,158$ users. We therefore exclude those users from the remaining users, leaving us with $89,603$ users to classify.

5.4 Step 2: Using basic profile information

Among the remaining $89,603$ users, in this step we attempt to use basic profile information to distinguish between minors and adults. Our heuristic is described below:

1. If a user provides his birth year , the user is adult.
2. If a user provides his high school name and graduation year, we say the user is adult.
3. If a user joined a high school network with graduation year is $<= 2009$, we say the user is adult.

4. If a user joined a high school network with graduation year is > 2009, we say the user is minor.
5. If a user has joined a college/grad school network or provided college/grad school graduation year, we say the user is adult.
6. If a user has joined a workplace network, we say the user is adult.

5.5 Step 3: Using Heuristics

After the above steps, there remains $55,376$ users. We use the following two heuristics to identify adult users and minor users.

1. Heuristic 1: With the default privacy settings for minors, public search will be disabled and only the "add him/her as a friend" message will appear in profile page. So if a user's public search status is disabled and only the "Add him/her as a friend" message appears in the public profile page, we classify that user as a minor. We classify $29,984$ users as minors in this way. There remains $25,392$ users to clasify.
2. Heuristic 2: In Section 3 and 4, we have estimated the age of all active users in NYC dataset. We apply the following heuristic to the remaining $25,392$ users: (a) If a user's estimated birth year is between 1930 and 1991, we say the user is an adult. (b) If a user's estimated birth year is greater than 1991 and his friend list is available, we say the user is an adult. (c) If a user's estimated birth year is greater than 1991 and his friend list is unavailable but his age can be estimated from reverse lookup, we say the user is a minor. (d) If a user's age cannot be estimated with iterations and his friend list is unavailable from profile page and reverse lookup, we say the user is minor. (e) If a user cannot be estimated with iterations and his friend list is unavailable from the profile page, but has less than 10 friends from reverse lookup, we say the user is a minor.

Combining all these steps together, among the 1.47 million NYC users, we classify 95.14% as adults, 1.40% as unknown and 3.46% as minors. Among the $19,488$ ground-truth minors, 95.45% were classified correctly; and among $1,081,567$ ground-truth adults, 100% we classified correctly.

6 Related work

We now review the prior work that considers inference of one or more private attributes in OSNs. To the best of our knowledge, this is the first paper that examines in-depth age estimation and minor classification in online social netowrks. Furthermore, our data set is at least one order of magnitude larger than all of those in the prior work on inference of private attributes (in the papers cited below).

Zheleva and Getoor [18] proposed techniques to predict the private attributes of users in four real-world datasets (including Facebook) using general relational classification and group-based classification. They looked at prediction of genders and political views, but not at age estimation or minor classification. Other authors [14, 17, 15, 13] have also attempted to infer private information inside social networks. Their methods are mainly based on link-based traditional Naive Bayes classifiers, and none of them consider the problem of age inference. Jernigan and Mistree [9] demonstrated a method for accurately predicting the sexual orientation of Facebook users by analysing 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.

Our work also relates to the problem of age estimation based on users' facial images as studied in [10–12]. In this class of work, the authors used publicly available aging databases (with facial images of users at different ages), and developed computer vision techniques for age estimation and evaluated their performance with respect to Mean Absolute Error (MAE). We achieve better results than these facial age estimation techniques using simple techniques that a naive attacker can easily exercise. Although we did not collect profile pictures of the Facebook users due to storage constraints, we note that profile images of Facebook users contain a lot of noise (e.g., due to low-resolution or lack of frontal view) and it would be hard to apply image-based age estimation for a large number of Facebook users. However, it would be interesting to combine our methodology and image-based techniques for further improvement of performance.

Becker and Chen [8] inferred many different attributes of Facebook users, including affiliation, age, country, degree of education, employer, high school name and grad year, political view, relationship status, university and zip code using the most popular attribute values of the user's friends. To our knowledge, this is the only other existing study that considers age estimation. Age estimation is not a focus of their study, and their dataset size has only 49 users. For this very limited study, their heuristics gave a success rate of 72.3%. In our paper, we examine a much larger dataset (over 49 million users) and develop a novel methodology that is based on limited profile information and on an interactive algorithm that exploits the underlying social network structure. We have applied our methods to a large data set of 1.47 million NYC users and verified on a set of 419 thousands ground-truth. Additionally, the minor classification problem is not considered in [8].

Mislove et al. [16] proposed a method of inferring user attributes by detecting communities in social networks, based on the observation that users with common attributes form dense communities. However, people with the same attributes, such as age and gender, may not form communities, and thus these attributes may not be accurately predicted using this approach.

7 Conclusion

In this paper, we investigated how difficult it is to estimate the age of OSN users who do not reveal their age publicly. To this end, we developed a novel two-step procedure. In the first step, we exploited side information such as high-school graduation year and high-school graduation years of friends with the same high school name to accurately estimate the age for a large set of users. In the second step, we exploited the underlying social network structure to derive an iterative algorithm, which derives age estimates based on friends' ages, friends of friends' ages, and so on. Our overall methodology is able to estimate age of 84% users in our dataset with a 4-year mean absolute error. However, we found that for many older users, age is difficult to estimate accurately, and may thus remain private within OSNs. We also developed a technique for another related privacy violation – classifying a user as a minor or as an adult. Our work casts serious doubts on age privacy and children online privacy in OSNs.

References

1. Cyber bullies harass teen on facebook before, after her suicide, available at: http://www.ktla.com/news/landing/ktla-facebook-suicide-bullies,0,133841.story
2. Developer blog: July 2009 platform news, available at: http://developers.facebook.com/blog/post/285
3. Facebook public search, available at: http://blog.facebook.com/blog.php?post=2963412130
4. Facebook statistics, available at: http://www.facebook.com/press/info.php?statistics
5. Facebook to fully deprecate regional networks by september 30, available at: http://www.insidefacebook.com/2009/08/05/facebook-to-fully-deprecate-regional-networks-by-september-30
6. How does privacy work for minors?, available at: http://www.facebook.com/help/?faq=16397
7. More cyberbullying on facebook, social sites than rest of web, available at: http://www.readwriteweb.com/archives/more_cyberbullying_on_facebook_social_sites_than_rest_of_web.php
8. Becker, J., Chen, H.: Measuring privacy risk in online social networks. In: W2SP (2009)
9. Carter Jernigan, B.F.M.: Gaydar: Facebook friendships expose sexual orientation. First Monday 14(10-5) (2009)
10. Geng, X., Zhou, Z.H., Zhang, Y., Li, G., Dai, H.: Learning from facial aging patterns for automatic age estimation. In: ACM Multimedia. pp. 307–316 (2006)
11. Guo, G., Fu, Y., Dyer, C.R., Huang, T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Transactions on Image Processing 17(7), 1178–1188 (2008)
12. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: CVPR. pp. 112–119 (2009)
13. He, J., Chu, W.W., Liu, Z.: Inferring privacy information from social networks. In: ISI. pp. 154–165 (2006)
14. 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)
15. Lindamood, J., Kantarcioglu, M.: Inferring Private Information Using Social Network Data. Tech. Rep. UTDCS-21-08 (2008)
16. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: Inferring user profiles in online social networks. In: WSDM (2010)
17. Xu, W., Zhou, X., Li, L.: Inferring Privacy Information via Social Relations. In: 24th International Conference on Data Engineering Workshop. pp. 154–165 (2008)
18. 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)