The City Privacy Attack: Combining Social Media and Public Records for Detailed Profiles of Adults and Children

Tehila Minkus
New York University
tehila@nyu.edu

Yuan Ding
New York University
dingyuan1987@gmail.com

Ratan Dey
New York University
ratan@nyu.edu

Keith W. Ross
NYU and NYU-Shanghai
keithwross@nyu.edu

ABSTRACT

Data brokers have traditionally collected data from businesses, government records, and other publicly available offline sources. While each data source may provide only a few elements about a person's activities, data brokers combine these elements to form a detailed, composite view of the consumer's life. The emergence of social media gives data brokers unprecedented opportunities to enhance their profiles. Data brokers are increasingly interested in combining the information collected from offline sources with information publicly available in social networks to profile not only adults but also children.

In this paper, we show how data brokers and other third parties can combine online and offline data sources – namely, public Facebook profiles and voter registration records – to create detailed profiles of adults, teens, and children in any target city in the US. We outline and execute an approach that leverages a Facebook user's social ties combined with the city's voter registration records to infer the Facebook users who reside in the city. These inferences enable a data broker to create detailed user profiles, which not only include information publicly available from Facebook but also the user's exact residential address, date and year of birth, and political affiliation. We further show how additional inferences can be made from the combined data. We then discuss how this city attack can be extended to create detailed profiles of minors and children. Finally, we make recommendations to Facebook, municipal authorities, and individuals to decrease the risk of this large-scale privacy breach.

1. INTRODUCTION

In a recent report, the Federal Trade Commission (FTC) alerted the public about the privacy risks of data brokers, calling for more transparency in the industry [2]. The primary business of data brokers is collecting personal information about individuals from a variety of sources in order to aggregate, analyze, and share that information and its derivatives for marketing products and performing credit checks. This information could also be sold to employers, dating sites, political parties, and college recruitment offices.

Data brokers have traditionally collected data from businesses, government records, and other offline sources. These include bankruptcy information, voting registration, consumer purchase data, warranty registrations, and other details of consumers' everyday interactions. While each data source may provide only a few elements about a consumer's activities, data brokers combine these elements to form a detailed composite of the consumer's life. Historically, data brokers have only offered lists about adults, rather than children or teenagers.

The emergence of social media gives data brokers unprecedented opportunities to enhance their profiles. Data brokers are increasingly interested in combining the information collected from offline sources with information publicly available online, such as social media and blogs [2]. Furthermore, because the children's market surpasses $200 billion in the US alone, it is not surprising that data brokers have recently also begun to compile dossiers on children as well [25] [26].

In this paper, we explore how profiles can be enriched by combining the public information available from a social network – namely, Facebook - with public records – namely, voter registration records. We consider not only the profiling of adults but also of minors, i.e. teens and children. We show that with just these two sources of information, data brokers can create alarmingly detailed profiles about adults, teens, and children. Data brokers with extensive financial resources and hundreds of employees can then build upon these profiles, creating a snowball effect.

1.1 Motivational Example

To motivate this study, let us consider an example of the potential outcomes from matching, say, an adult woman listed in a voter registration record and a specific Facebook account. Suppose that this woman has two children living at the same address, one son in high school and one daughter in elementary school. Further suppose the mother posts pictures of her children to Facebook, and her privacy settings make these photos publicly available [18]. Finally, we can also make the very reasonable assumption (as discussed in the body of this paper) that the voter record for this adult contains the woman's exact home address, birth date and year, and political affiliation. Then by combining this woman's Facebook profile with her voter registration record, the data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

COSN'15, November 2–3, 2015, Palo Alto, California, USA.
© 2015 ACM. ISBN 978-1-4503-3951-3/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2817946.2817957.

broker immediately knows the woman's exact street address, birth day and year, political affiliation, profile picture, and all her public Facebook profile information, possibly including additional photos, her list of Facebook friends, education, workplace, likes, and so on. Beginning with this profile, the data broker can potentially further infer gender, religion, sexual orientation, economic level (as based on address of residence), ethnicity, and so on [16] [27] [6].

Continuing with the example, the data broker can also use Facebook to find all the high-school students living in the the adult woman's city [8]. For each of these students, the data broker can create profiles that include name, gender, profile photo, high school, graduation year, and friends. By matching the last name of the woman with the last names of the high-school students (or by making use of Facebook friend lists), the data broker can identify the Facebook account of the woman's high-school son. The data broker can then enhance the profiles of this high-school student by combining the high-school profiles in [8] with the information in the voter records, including exact home address, parents' full names, birth dates, and political affiliations.

Since the mother is also posting pictures of her children, the data broker can also compile dossiers on young children who may not even have Facebook accounts [18]. These profiles can include the child's name, birth date, and home address, as well as his parents' full names, birth dates and political affiliations.

1.2 Beyond Data Brokers

Up until this point, we have focused our discussion on profiles created by data brokers. But other parties are potentially interested in combining public records with social media to obtain combinations of social, personal and demographic datapoints. For example, as the political industry refines microtargeting techniques [15], there is increased demand for fine-grained information about individual voters. By matching a voter registration record to a specific Facebook profile, a political campaign would be able to send personalized messages to the voter in order to sway or influence his voting behaviors. However, a 2012 survey of American voters found that a majority of voters responded negatively to the idea of political data collection and microtargeting [28]. Moreover, creating a list of registered voters on Facebook would also identify which Facebook users are not registered to vote. This would allow political campaigns to target non-voters in voter registration drive for get-out-the-vote purposes.

The profiles could also be used to fuel a large-scale and highly personalized spear-phishing attacks. Messages could automatically be generated to include the target's address, birth date, Facebook friends and so on, with the goal of tricking the targets into installing malware or providing financial information to enable financial cybercrimes.

Finally, consider Facebook's own potential interest in public records. Because Facebook earns virtually all of its revenues from targeted advertising, there is a direct correlation between Facebook's stock price and the amount and quality of the information it has about its users. Facebook, as well as other social networks, is likely interested in obtaining offline information from public records (or indirectly from data brokers), and combining this information with the data it directly obtains and infers from user activity on Facebook. In fact, Facebook as already partnered with Acxiom, one of the largest data brokers [3].

1.3 Contribution

Combining information from Facebook and voter registration records hinges on *the ability to match a person in a voter registration list with a Facebook user with a high degree of certainty.* The problem is challenging because a name in a voter registration list can match with hundreds of Facebook accounts. In this paper, we show how a data broker can simplify the task by performing the voter-to-Facebook user matching on a *city-by-city basis.* Within a city, the number of possible name matches is substantially reduced, often to only one. When there are multiple matches, additional information in Facebook and the voter records can often reduce ambiguities. We refer to the *City Privacy Attack* as the problem of attempting to profile all the residents of a given city – including adults, teens, and children – by matching voter records with Facebook users.

In principle, for a given target city, matching the Facebook users residing in the city to the adults in the city's voter list should simplify the problem. But most Facebook users do not provide their current city information in their public profiles, making it difficult to match them to the voter records. In this paper, we develop a novel methodology for inferring the Facebook users who live in the target city. This methodology combines the information in the voter lists with the public Facebook friend lists. As a case study, we select a small city in the Northeast USA for analysis. After obtaining the voter list from municipal authorities, we propose and evaluate several methods to match each voter to a single Facebook account. We also discuss how the attack can be extended to profiling the minors and children in the city, show how the profiles can be enhanced by inference techniques, and finally make recommendations to Facebook and about voter data use.

The structure of this paper is as follows. In Section 2, we discuss voter registration records and their prevalence throughout the United States. We also detail how the characteristics of our dataset. In Section 3, we consider several naïve approaches to matching voter data with Facebook profiles. Since these approaches prove unsuccessful, we introduce a more sophisticated approach in Section 4 using social ties to match Facebook users to voters from the targeted city. In Section 5, we analyze the results and limitations of this approach. In Section 6, we show how this attack can be extended to teens and children in the targeted city. In Section 7, we show how the combination of voter and Facebook data allows an observer to infer new traits about the targeted voters. In Section 8, we present a survey of related work, and in Section 9 we discuss the implications of our findings. Finally, in Section 10, we conclude.

1.4 Ethical Considerations

Conducting privacy research on public data can be ethically sensitive. In this work, we took measures to ensure minimal risk of exposure for any individuals whose data was studied. For this reason, we do not mention identifying details about individuals in the course of this paper. Moreover, we believe that it is important to discuss the privacy risks that can attend public release of data, and that the benefits of public discourse on this subject outweigh the risks.

2. VOTER REGISTRATION RECORDS

According to a study by the California Voter Foundation in 2004 [4], all 50 states in the USA require voters to pro-

Data field	States Collecting
Name	50
Address	50
Signature	50
Date of birth	49
Phone number	46
Gender	34
All or part of SSN	30
Party affiliation	27
Place of birth	14
Driver's license number	11
Race	9
Special assistance requirements	4
Parent's name	3
Email address	2
Occupation	1

Table 1: Information that different states require voters to supply [4].

Data field	States Redacting	Sharing
Birthdates (some or all)	11	38
Phone numbers	5	41
Social Security numbers	29	1
Birthplace	2	12
Driver's license numbers	6	5

Table 2: Among the states who collect certain data fields, the number of states who redact the information or share it with third parties [4].

vide their name, address and signature prior to voting. In addition, many mandate that voters must additional information, such as phone number, gender, Social Security number, and additional demographic data. For a breakdown of state requirements, see Table 1.

These data are collected under the auspices of voter registration, yet their use is not strictly confined to usage by voting registrars and poll workers. The lists are put to use by government and judiciaries: in all 50 states, political parties and candidates are granted access to the voter rolls, and 43 states use voter lists as a source for jury duty service. 27 of the states allow certain voters (such as public figures or victims of domestic violence) the right to retract parts or all of their voter registration records before the lists are shared.

Beyond these political and judicial uses, voter data is shared with third parties in many states. 22 states grant unrestricted commercial access to their voter rolls. While some states redact certain fields (see Table 2), many data fields are left visible to anyone who can access the voter registration lists.

2.1 Obtaining Voter Data

For the analyses in this paper, we arbitrarily selected a small city in the Northeast USA. As per the 2010 United States Census, the target township has approximately 70,000 people, 27,000 households, and 19,000 families.[1] The city is approximately 78% white and 22% minorities. The median annual income is about USD 90,000.

After contacting the municipality, we were able to purchase a CD of the voter registration records for USD 35.00. The records obtained include all adult citizens who have registered to vote in the specific county/municipality (check out terminology). Each voter registration record contains the following fields:

- **Name:** Every voter's first and last name was included. Additional fields, such as middle name, prefix, and suffix, were optional.
- **Gender:** 17.3% of the voters were male, 20.1% were female, and 62.5% chose not to identify their gender. (In Section 7.1 we discuss methods to infer genders for these voters.)
- **Address:** Each voter's complete street address and zip code was included.
- **Political affiliation:** Voters were allowed to choose between Democrat, Republican, and unaffiliated. 38.1% chose Democrat, 16.1% chose Republican, and 45.7% are not officially affiliated with any party.
- **Date of birth**: For all voters, the month, day, and year of birth were included.
- **Other dates:** The date on which each voter registered to vote is also included, as well as the date upon which they officially received their voting privileges. Additionally, when applicable, the date of the voter's party registration is also indicated.

Notably, the voter registration records do not specify the user's family members or marital status. In some cases, the voter's gender is also missing. In Section 7, we introduce methods to infer these characteristics as well.

3. TARGETED-INDIVIDUAL APPROACH

There are a variety of ways to attempt matching voter records to Facebook profiles. In this section, we discuss the category of approaches that focus on searching Facebook for specific voters, one by one.

3.1 Matching Voters by Name

Facebook uses a real name policy, requiring users to supply their actual names when they register with the social network. As such, it would seem easy to match a voter's name to a Facebook user's name. But this simple process is flawed due to the frequency of common names in Facebook. We illustrate with a small experiment, where we randomly choose 100 names from the voter registration list and search those for those names in the Facebook name directory.

For each user, we inserted their name into the custom search URL `https://www.facebook.com/search/str/NAME/users-named/intersect` to search for matching profiles, replacing the string `NAME` with the user's first and last name. We then counted the number of profiles returned by the search.

In Figure 1, we show the distribution of how many people were matched to no profiles, one profile, or multiple profiles when using their names as keywords. As the figure shows, this results bears some strong limitations. A distinct, one-to-one Facebook profile match was found for only 6% of the

[1] A family consists of two or more people (one of whom is the householder) related by birth, marriage, or adoption residing in the same housing unit. A household consists of all people who occupy a housing unit regardless of relationship[1]

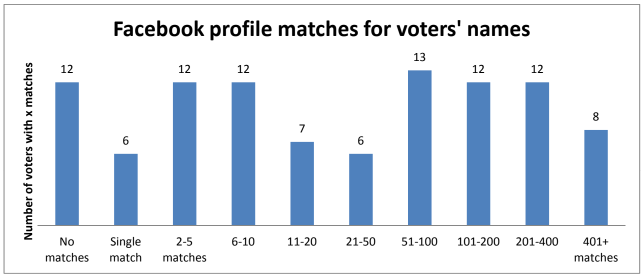

Figure 1: Matching voters by name across all of Facebook.

randomly selected voters. More than half of the selected voters (58%) had more than 10 Facebook profiles that share their names. Therefore, it becomes apparent that this approach casts too wide a net. Since more than 1 billion people use Facebook[2], even relatively rare names can occur several times in the Facebook population. As such, using name as the single criterion is insufficient to match voter records to Facebook profiles.

3.2 Geo-targeted approach

What if one employs the location information from a Facebook user's profile to help narrow down the search space? Using location information can increase the matching precision, but many Facebook users do not include their current city in their public profile. For example, Dey et al. [9] found that only 36% people provided current city information publicly. As such, relying on the users' explicit location information will overlook many users who have not shared their location in their public profiles.

To illustrate this, we also searched for the 100 randomly selected voters on Facebook along with their city. We used the URL `https://www.facebook.com/search/str/NAME/users-named/intersect/str/CITY-ID/residents/intersect`, replacing `NAME` with the user's name and `CITY-ID` with the numerical ID of the targeted city. Of the 100 voters, 82 had no Facebook matches at all. 16 voters had exactly one match, while one voter had two matches and another voter had three matches.

As this experiment shows, searching for users' names and explicit location data has very low recall. Additionally, this class of approaches carries a high overhead, since an individual query is required for every user. In the next section, we explore a more efficient and effective approach towards matching voter records with Facebook profiles.

4. CITY ATTACK

In the previous section, we used the names from voter records as keywords to search for corresponding Facebook profiles with limited success. In this section, we develop a novel methodology for inferring the Facebook users who live in the target city. This methodology combines the information in the voter lists with the publicly available friend lists.

Figure 2 summarizes the approach. Using Facebook graph search, we first find some of the people who live in the target city. Specifically, after logging into Facebook, we enter the URL `https://www.facebook.com/search/str/CITY-ID/residents/intersect` in the address bar of the browser, replacing `CITY-ID` with the numerical Facebook ID of the city. Facebook returns a list of some of the people who live or have lived in the target city; this list auto-populates as the user scrolls down. We automate this process, continuing to browse until we find "End of Results". This provides a partial list of people (including their Facebook names and IDs) who are currently living in the target city. We then add their IDs to a seed list as shown in Figure 2. For our target city example, we automated this process with four accounts, yielding a list of 10,200 people who currently live in the target city.

We then attempt to find corresponding matches in the voter registration records, using the first and last names from the Facebook profiles. We also enriched these lists using an auxiliary database of common nicknames. For each Facebook page, we checked if the first name was included in a list of common nicknames. If it was, then we also tried to find voter record matches for the full names corresponding to that nickname. We then put these matched IDs in the "match pool". Of the 10,200 seed Facebook accounts, we found corresponding voter records for 5,294. We then proceed as follows:

1. Repeat for all members of the match pool (until some threshold is reached):
   (a) Retrieve the user's friends list.
   (b) Check if a friend's name is found in the voter list. If so, add the friend to the potential match pool.
2. Output the match pool as the list of potential matches.

If a user self-identifies as a resident of the target city and also matches to a voter record, we consider the user to be a potential match. For Facebook users who have not identified their location, there is reasonable likelihood of being registered voters in the target city, since both *their names* and the name of *at least one respective friend* are found in the voter registration list.

5. RESULTS

Since there is no ground truth readily available for this dataset, we instead measuer the overall accuracy of the matches by employing a set of heuristics and filtering results. We report our results in this section.

Following the approach detailed in Section 4, we used Facebook's Graph Search to download a list of users in our target city. Beginning with these seeds, we iteratively crawled over 1 million users and identified 35,556 Facebook accounts whose name or nicknames correspond to an entry in the voter registration list.

For some of the voters, more than one match was found, since multiple Facebook accounts with the voter's name were discovered over the course of the crawling process. Thus, the set of candidate-match Facebook accounts corresponded to a smaller set of voter records. Specifically, 20,481 voter registration records had one or more matches in the Facebook set after two iterations. We stopped at this point since we had reached a significant coverage of the voting registration list (37 percent of records).

5.1 Filtering for Accuracy

[2] https://newsroom.fb.com/company-info/

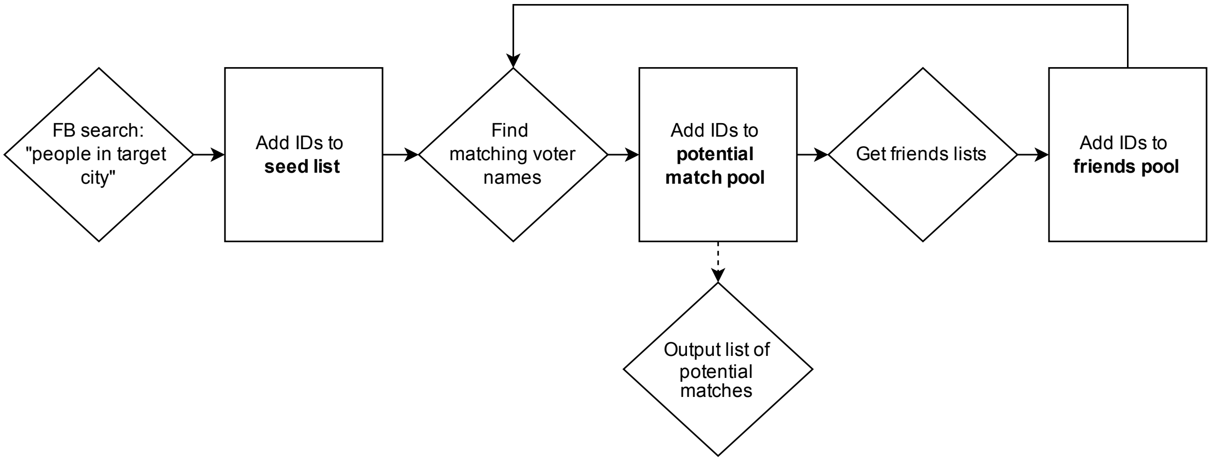
FB search: "people in target city"
Add IDs to **seed list**
Find matching voter names
Add IDs to **potential match pool**
Get friends lists
Add IDs to **friends pool**
Output list of potential matches

Figure 2: The process for crawling potential residents of the target city.

In total, we had 20,481 records in the voter registration list matching to 35,556 users in Facebook. Among these Facebook users, we found that 15,501 of them had entered their current city and hometown as places other than the target city; therefore, we exclude them from further analysis. After this filtering step, we have a candidate set of 20,055 users in the set of candidate Facebook matches, corresponding to 16,457 voter records.

Among these users, 7,197 have shared on Facebook that their hometown or current location is the target city. The rest of the users have not specified where they live, but they have an average of approximately seven friends who have shared that they live in the target city. Considering the relatively small size of the target city, combined with the fact that many users don't share their location, this indicates that these users with unknown locations are likely to live in the target city.

5.2 Analysis of Social Ties for Location Inference

To ensure correct inference of the users' location, we conduct another filtering step based on the users' friends lists. Intuitively, we can expect that if a person has many friends from a given location, he is likely to be from that location as well.

For this purpose, we introduce a measure denoted f. For a user who has a public friends list, we can count how many friends he has among the other potential residents (in our case, 20,055 users); so for a Facebook user i, f_i is the number of friends in his list who are also in the potential match pool set.

5.2.1 *Reverse Friends-List Lookup*

Among the 20,055 Facebook candidate matches, 8,283 users hid their friends list from the public. This would make it difficult to measure the strength of their social ties within our dataset. However, the design of the Facebook friends list enables one to learn about a user's friendship ties based on the information that others have made public [8].

For example, imagine that John and Martha are both users of Facebook. John is privacy-conscious, so he hides his friends-list. Martha is more interested in a robust online social life, so she shares her friends list. Since Martha is friends with John, we are now able to learn of at least one of John's friendships even though he hid his friends list. If all of John's friends share their friends lists, then we will be able to learn all of his social ties despite his restrictive privacy settings.

We leverage this reverse look-up method to introduce an alternate measure for f for users who have hidden their friends lists: we count how many other users in the dataset have listed them as friends (i.e. for user i, $f_i = x$ if we can find this user in the friends lists of x other users in this dataset). In this way, we can indirectly infer how strongly they are socially tied to the other members of our dataset.

This method is not guaranteed to retrieve all friendships of the users with private friendslists; therefore, the inferred friend lists are usually shorter than the public friend lists. To correct for this disparity, we introduce a corrective weighting scheme. On average, the users with public friends lists had 189% as many friends in the dataset as those who did not share their friends lists. Therefore, we multiply the friend-count of the users who hid their friendslists by a factor of 1.89. This ensures that the f-measure of the more private users is on the same scale as the other users.

5.2.2 *Parameter Selection*

If user i has a high value for f_i, i.e. he has many friends who are in the dataset, then it is likely that he is indeed located in the target city. For lower values of f, more potential matches are allowed; higher values of f restrict the set of matches to users who have more social connections within the potential match pool. A natural heuristic is to assume that a Facebook user i is a resident of the target city if either (i) the user self-identifies as a resident of the city; (ii) or has an f_i value at least equal to some threshold f.

Let $N(f)$ be the number of Facebook users in the possible match pool who have f_i values of at least f or say they live in the target city. Thus $N(1) =$ 20,055. Figure 3 plots N(f) as a function of the threshold F. For lower values of f, there are more matches allowed. As f is raised, the matching criteria become more restrictive and allow for fewer matches.

While choosing lower values of f allows for more matches, these matches are less precise and more prone to duplicate matches for a single voter. Raising the f-value results in fewer duplicate matches. Thus, fewer voters are matched

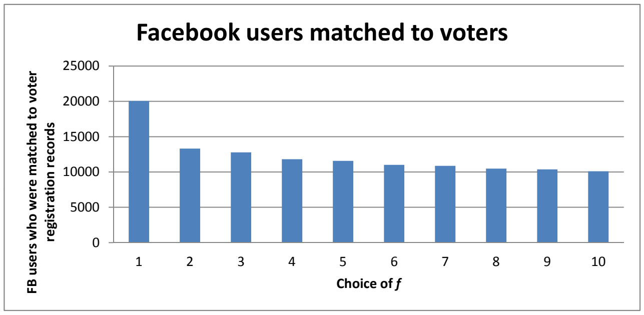

Figure 3: The number of Facebook users who were matched to a voter registration record, as a function of the choice of f.

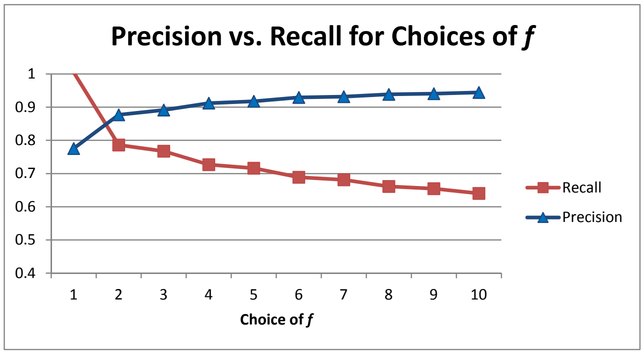

Figure 4: Precision and recall for each value of f. Precision measures the percentage of matches that were one-to-one. Recall measures the percentage of voters in our filtered set who were matched to at least one Facebook profile.

respectively to multiple Facebook pages, increasing precision. But raising the choice for this parameter also decreases the overall recall, by allowing fewer potential matches to be considered. As such, setting the f-value is a tradeoff between precision and recall. We quantify this tradeoff in Figure 4. Precision for a given value of f refers to the percentage of one-to-one matches among all users in $N(f)$. Recall for a given value of f refers to the percentage of the 16,457 voters in our dataset who had a match suggested from among $N(f)$.

5.3 Limitations

We point out several limitations which make it difficult or impossible to determine a precise match for each of the voters in the registration list. Firstly, it is likely that not all voters are registered users on Facebook. According to a Pew Research poll in September 2014, only 58% of American adults aged 18+ used Facebook [11]. For these voters, there are no corresponding Facebook accounts that can be correctly matched to their records.

Secondly, there are name-related ambiguities which can be difficult to solve. We explore possible reasons for finding multiple matching Facebook accounts for a single voter record. We randomly select 50 accounts from voting list who have double matches among the Facebook accounts, and discovered three patterns:

- **Multiple account creation**: Some people may retain multiple Facebook accounts. For example, two Facebook accounts matching to the same voter name shared a name, location, occupation, and some contacts, but one of the accounts has no new posts since 2012. It seems likely that this user abandoned the older Facebook account in favor of a new account but never deleted the older account. It would be of interest to attempt to filter out the duplicate inactive accounts, although it is not pursued here.

- **Multiple people with the same name**: There might be two or more people in the city with the same name, with only one of them registered to vote. For example, one voting record in our dataset was matched to two Facebook pages that seemed to belong to different people. As such, it is likely that for some names, there are multiple residents in the target city yet not all of them have registered to vote.

- **Low local-friendship threshold (f-measure)**: If a low threshold is set for the measure of local friendships, then incorrect matches might be suggested for voters. For example, in our dataset, one voter record matched to two Facebook accounts; one of these Facebook accounts had 26 local friends, but the other one has only 3 local friends. Thus the first match is more likely to be correct than the second.

Finally, when multiple voters share the same name, they may all be matched to the same set of Facebook accounts. For example, if several voters in the town are named Jane Doe, and there are also many women in the candidate Facebook set who are named Jane Doe, then there will be many potential matches between the two datasets. To disambiguate these cases, we may be able to use ages. Recall that the voter records typically give the ages of the voters. Although most Facebook users do not make their age publicly available, following the work of Dey et al. [10], it may be possible to estimate the ages of the Facebook users and thus employ age as a factor for more precise matching between voter records and Facebook. Additionally, by using face recognition and age estimation software, it may be possible to estimate a Facebook user's age based on his profile photo [13]. This estimated age could then be used as another data point in finding a correspondence to a distinct voter registration record. We leave this approach to future work.

We emphasize that the intention of this work is to provide a proof-of-concept for the idea of matching between offline databases and online social networking account. Since we did not have ground truth for this dataset, we have instead relied on reasonable assumptions to guide our algorithms towards likely matches. By demonstrating several matching techniques, we introduce a lower bound on matching accuracy and recall for real-life and online databases.

6. PROFILING TEENS AND CHILDREN

In the previous sections, we detailed how the combination of voter records and Facebook profiles enables an attacker to profile a large portion of the adults in a city. We now describe how it is possible to extend this profiling attack to high school students, who may have Facebook accounts but are too young to register to vote, as well as to children, who do not even have their own accounts on Facebook.

6.1 Profiling Teens

Facebook takes precautions to limit third parties from using its services to discover and profile minors. These precautions include banning young children from joining, excluding minors from search results, and displaying minimal public information for minors, no matter how they configure their privacy settings. Dey et al. recently showed, however, that an attacker, with modest crawling and computational resources, and employing data mining heuristics, can circumvent these precautions and create extensive profiles of *most* of the high school students in any targeted city [8]. Since some children lie about their ages when registering, this increases the exposure for themselves and also for their non-lying friends. In particular, using Facebook and for a given target high school, the attack described in [8] finds most of the students in the school, and for each discovered student infers a profile that includes significantly more information than their initial public profile. The information minimally includes the student's full name, profile picture, current high-school, graduation year, inferred birth year, and list of school friends.

We now outline how it is possible to significantly enhance these teen profiles by leveraging the techniques in this paper. Specifically, for each high-school student discovered and profiled using the techniques in [8], we can attempt to match the student to his/her parents in the voter registration list. There are two natural approaches to perform this matching. The first approach is to simply match the student's last name to the last names in the voter registration list. Given that most children inherit the last name of one or more of their parents, this simple approach should provide matches for almost all the high school students discovered in [8]. However, for students with common last names, there will likely be multiple matches, resulting in some ambiguity.

The second approach exploits the fact that some parents are Facebook friends with their children. In this approach, we examine the friend lists of all the high school students and also the friend lists of all of the adults profiled by the techniques in this paper, and look for common last names between adult and high-school student. As compared to the first approach, this last approach will give fewer duplicate matches but will also not provide as much coverage, as some children are not Facebook friends with their parents.

Once a high-school student is matched with a parent, then in addition to the profile information obtained in [8], the attacker now knows the teen's exact home address as well as the parent's full name, birthdate, marital status (see Section 7), and political affiliations. This additional information – and in particular the exact home address – is particular sensitive and can even put the teens at risk.

6.2 Profiling Children

Minkus et al. [18] showed that many parents post their children's photos along with their names and birthdates. This allows an outside observer, online service provider, or surveillant authority to learn facts about these young children. Using the techniques in this paper, a parent's Facebook account can also be matched to a voter registration record, allowing the attacker to develop detailed profiles of the parent. Thus, when a parent posts photos of her child, an attacker can generate a profile of the child that includes the child's address, name, birthdate and photos, as well as everything obtained about the child's parents using the methodology described in the previous sections. This is troubling from a privacy perspective, particularly since these children are often too young to maintain their own Facebook accounts or consent to their information being shared.

7. ENHANCING THE COMBINED DATASET

In Section 4, we developed a methodology for matching Facebook users to voter registration records. The combination of these data sources allows a third party to construct profiles based on the voter records as well as social and personal traits gleaned from the public Facebook page. However, many important data fields are still sparse or unavailable in this dataset. Specifically, many of the voters and Facebook users have not specified their sex, marital status, or family units. These traits are of interest to third-party data aggregators, such as data brokers, political parties, and advertisers. Can they be inferred by using the combination of Facebook profiles, voter records, and some auxiliary information from the public domain?

7.1 Filling in Missing Genders

In the voter registration records that we acquired, many voters had not provided their gender; specifically, 62.5% of the registered voters had left the gender field blank. In this section, we detail an approach towards inferring a person's gender based on his or her name and some auxiliary data. Our approach enables us to assign a presumed gender to 91.9% of all the voters.

We collected a list of auxiliary name data. The United States Social Security Administration makes available a list of the top 1,000 names for boys and girls born in specific years or decades[3]. We collected and parsed the lists spanning the decades from 1910 through 1999, which encompassed the birth years of the voters in our targeted city. We created two name directories, one for boy names and one for girl names, with each name weighted by the number of times it had appeared in the top-thousand list for that gender.

We then used this auxiliary data source to predict a gender for voters who had not specified one. If a voter's name appeared only in the SSA list of boy names, we presume that voter to be male; likewise, if the name appears only the in the SSA female name list, then we consider that voter to be female.

However, in some cases, a name appeared in both the male and female lists of the SSA, thus leading to some ambiguity. For example, the name Linda was a top-thousand name for girls in all of the nine decades under consideration, but in three of the decades in question, it was also a top-thousand name for boys. In such cases, we apply a simple heuristic: since the name Linda appears more often in the top-thousand lists as a girl name, we assume that a voter named Linda is a woman. Using the SSA records with these heuristics allowed us to learn the genders of an additional 27,905 voters (83.9% of the voters who had not provided a gender).

Finally, some names were not included in the SSA lists due to their relative obscurity. To provide better coverage of these names, we utilized the voter registration lists to find other voters who had specified both such a name and the given gender. In cases where the name had been used for both men and women in the voter registration records, we resolved it by simple voting as explained in the previous

[3] Available at `http://www.behindthename.com/top/`

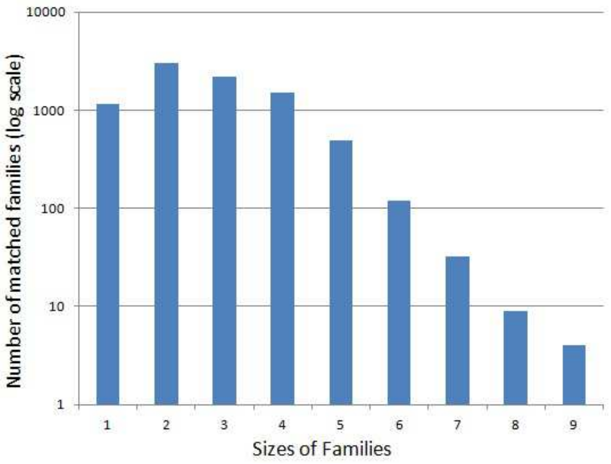

Figure 5: The number of families of each size, among families where at least one member who was matched to the voter records.

paragraph. This technique enabled us to learn the names of an additional 733 voters, or 14.5% of the voters who had not specified a gender and whose names did not appear on the SSA top-thousand lists.

After adding the 20,097 voters who initially provided their gender (37.5% overall) to the 29,178 voters for whom we inferred a gender (54.4% overall), we are left with gender data for 91.9% of voters.

7.2 Households and Family Units

We utilize the home address information of the 53,600 registered voters to infer households and family units. We predict household units using a simple heuristic: if people live in the same home or apartment, then we consider them to be a family or household. Specifically, if the house number, street name, and residential unit are identical for two or more residents, we consider them as a household. Using these approach, we group the voters into 25,007 households.

For each household, we check if any members have been precisely matched to a Facebook page. 8581 households have at least one member who was matched. Most of these households have only one member who has matched with the Facebook list; less than 10 households have six or more members who were matched to Facebook profiles.

In Figure 5, we show the distribution of family sizes for these 8,581 households. This family size is based on the voter registration list. Most of the households have 2-5 members, which is typical in the US.

7.3 Determining Marital Status

In this section, we examine the 25,007 presumed families to discover married couples. We use the following heuristics:

- Suppose there are only two people in a household, one male and one female, both of them are over 21 years old. Then we say they are married.

- Suppose there are three or more people in a household. Among those in the household, suppose there are a man and a woman, both of them are over 25, and the age difference between them is less than ten. Then we say they are married.

- All other people are presumed to be single.

After applying the above heuristics to our voter registration list augmented with gender predictions (as detailed in Section 7.1), we can detect 12,732 married couples among the 25,007 households.

Table 3 shows which traits can be learned about the voters by using each of the data sources individually, or in combination, or in combination with inference. Personal attributes like political affiliation, birth year, date and full home address are included in voter registration records. In Facebook, more than 95% of the people had hidden these fields from strangers, yet combining their profiles with voter registration data (when available) exposed this information.

7.4 Further Inferences

In this paper, we showed that the combination of voter records and Facebook profiles can be used to infer new facts about voters, such as gender, marital status, and family connections. In addition, it would be possible to infer additional facts based on the social connections and profile data of the users in question. For example, a third party could enrich these profiles with sexual orientation and race or ethnicity. This brings privacy concerns about big-data techniques into stronger relief.

Sexual Orientation.

Jernigan and Mistree [16] also showed that one can leverage a Facebook user's friends list to infer his or her sexual orientation. They found that the proportion of gay friends that a user had on Facebook held predictive power for classifying the user as straight or gay. This held even when a user had hidden or omitted his sexual orientation on his Facebook profile. Thus, the combination of these two data sources would also enable relatively accurate prediction of voters' sexual orientations, even though the majority of users did not specify their sexual orientations on their profiles.

Race and Ethnicity.

Elliott et al. [12] showed that by using U.S. Census records of popular last names, one can predict a given person's ethnicity with high accuracy given his last name and zip code. They implement a Bayesian approach using the prior probabilities of racial identity given a specific neighborhood, and they update this with the probability of the person's specific last name belonging to that ethnicity. Since our dataset includes each voter's last name as well as their address, one would be able to apply this method to our dataset to output predictions for each voter with regard to race and ethnicity.

8. RELATED WORK

8.1 Combining Online Profiles

A considerable body of research has explored methods for associating accounts from multiple online services to individual users. Notably, Narayanan and Shmatikov showed that the Netflix prize data, though anonymized, could be matched to the Internet Movies Database (IMDB) website to identify users [20]. Irani et al. [14] showed that by participating in larger numbers of online social networks, users exposed more private information. More recently, Perito et al. [21] showed that users could be linked across online services through their username choices. For example, Minkus and Ross [19] showed

Data field	Facebook	Voter registrations	Combined with inferences
Political affiliation	0.002%	100%	100%
Birth year (and age)	0.02%	100%	100%
Birth date	0.04%	100%	100%
Physical address	0%	100%	100%
Sex (M/F)	16.6%	33.2%	93.9%
Education	49%	0%	49%
Religion	0.003%	0%	0.003%
Sexual orientation	9%	0%	9%
Relationship status	21.1%	0%	56.6%
Friends list	21.2%	0%	96.3%

Table 3: For all users with a distinct match between a Facebook profile and a voter record, we show which data fields were available from the Facebook profile, from the voter record, and from the combined profile when enriched with inferences based on auxiliary data.

that eBay accounts could be matched to Facebook profiles, thus revealing a user's real name sand purchase history.

More generally, a large body of work has examined the problem entity matching in databases; see Köpcke and Rahm [17] for a survey of notable approaches. Entity matching deals with the problem of finding and resolving duplicate records across multiple databases. While this is similar to our problem, we find that our specialized approach is better able to leverage the unique properties of our datasets, thus vividly demonstrating the privacy risks of record linkage between social networking sites and public data.

8.2 Relating Online Data to Offline Data

In this paper we explored in some detail how to connect Facebook data with voter registration lists. Barbera [5] matched the voter registration records of Ohio voters to Twitter accounts, using their full names and counties and filtering out any duplicate matches. However, only users who explicitly provided their location were analyzed; no attempts were made to infer the location of users who had not included it in their profile. Chen et al. [7] used social media profiles to identify users' phone numbers and addresses based on online phonebook records, also removing any duplicate matches. However, their work fails to account for the many users who have unlisted phone numbers or cell phones that are not listed in phonebooks. Additionally, while phonebook records include address information, they do not include birthdates or political affiliation, which are contained in voter registration records. Finally, in both of these works, no methods were proposed to resolve ambiguous matches between multiple accounts. Our work is more rigorous in attempting to resolve ambiguous matches between multiple profiles and voter records. We utilize social relationships as side-channel clues that can hint at a user's undisclosed location, thus providing an additional datapoint that can narrow down the matching process.

Some work has also focused on learning the physical locations of users based on their social media activity [22] [23]. However, our approach does not rely on geo-tagged data or posts that are tied to specific locations. This extends the attack's coverage to users who do not have geo-tagged or location-specific data included in their posts.

9. DISCUSSION

In the previous sections, we introduced a technique for automatically matching large numbers of registered voters to Facebook profiles. We also leveraged the combined dataset to infer new descriptions of the targeted voters. In this section, we discuss what can be done to limit the privacy risks to individual voters who use Facebook.

9.1 Recommendations to Facebook

The attacks described in this paper were enabled primarily by two Facebook policies: the *real-name policy* and the *reverse-lookup friends lists capability.* We describe how changes to these two policies would allow users to exercise better control of their private information.

Abolish real names policy.

According to Facebook's Help Center[4], "Facebook is a community where people use their authentic identities. We require people to provide the name they use in real life; that way, you always know who you're connecting with. This helps keep our community safe... Pretending to be anything or anyone isn't allowed." While this policy is articulated in a manner that emphasizes trust, it also means that users may not use false names as a way to protect their privacy.

By abolishing the real-name policy, Facebook would allow users to create false identities to hide from advertisers, data brokers, mass surveillance, and unwelcome snooping by social acquaintances. The social network Google+ recently changed its real-name in order to be more inclusive, since they felt it "excluded a number of people who wanted to be part of it without using their real names". We recommend that Facebook do the same.

Implement symmetric privacy settings.

In the current implementation of the Facebook privacy settings, a good deal of information can be leaked by a private user's friends. Specifically, in this paper we leveraged the *reverse-lookup* capacities of the friendship list mechanism. Even if a user hides his own friends list, he can be found on the public friends list of any one of his acquaintances. Due to the social property of homophily, this often reveals a good deal about a user beyond his social ties; for example, gay people are more likely to have a higher proportion of gay friends [16], and a user's age can also be estimated by analyzing the ages of his social ties [10].

For these reasons, we recommend that Facebook institutes a *symmetric privacy policy* with regard to friendship ties.

[4] `https://www.facebook.com/help/112146705538576`

Namely, we believe that a user who hides his friends list should also be removed from the public friends lists of any other users. Alternatively, Facebook may consider making all friend lists invisible, even to friends. We note that WeChat, the hugely popular social network in China, has made this design choice, and is therefore not plagued like Facebook with inference attacks based on friend lists.

9.2 Voter Data Use Recommendations

As political campaigns increasingly digitize, the use of campaign and voter data in both political and commercial uses is gaining traction [15]. In a 2014 paper, Rubinstein [24] made several recommendations for better privacy practices governing the collection and use of voter data. We summarize them here:

- Increased transparency: when collecting data about voters, authorities should clearly state any uses for which the data may be used. Moreover, any secondary uses of the data should include a disclaimer about the origins of the voter data.
- Restricting commercial data practices: similar to the Fair Credit Reporting Act, individuals should be given the right to correct, remove, or access any of the information about them that is sold commercially.

10. CONCLUSION

In this paper, we empirically examined the problem of matching online and offline profiles. Specifically, we showed that matching voter records to Facebook profiles, though difficult, can be accomplished by leveraging both explicit and implicit features of the datasets. We then showed that the combination of these data sources allowed for new, richer inferences with negative privacy implications. Finally, we suggested policy changes to better protect the privacy of Facebook users and voters.

Acknowledgements

This work was supported in part by the NSF (under grants CNS-1318659 and DGE-0966187). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any of the sponsors.

11. REFERENCES

[1] Frequently asked questions. United States Census Bureau. Available: `https://www.census.gov/hhes/www/income/about/faqs.html`.

[2] Data brokers: a call for transparency and accountability. *Federal Trade Commission, Washington, DC*, 2014.

[3] Acxiom becomes an audience data provider in facebook marketing partner program. February 17, 2015.

[4] K. Alexander and K. Mills. Voter privacy in the digital age. *California Voter Foundation*, 2004.

[5] P. Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91, 2015.

[6] J. Chang, I. Rosenn, L. Backstrom, and C. Marlow. epluribus: Ethnicity on social networks. *ICWSM*, 10:18–25, 2010.

[7] T. Chen, M. A. Kaafar, A. Friedman, and R. Boreli. Is more always merrier?: a deep dive into online social footprints. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pages 67–72. ACM, 2012.

[8] R. Dey, Y. Ding, and K. Ross. The high-school profiling attack: How online privacy laws can actually increase minors' risk. In *Proc. of Internet Measurement Conference*, volume 13, 2013.

[9] R. Dey, Z. Jelveh, and K. W. Ross. Facebook users have become much more private: A large-scale study. In *PerCom Workshops*, pages 346–352. IEEE, 2012.

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

[11] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, and M. Madden. Social media update 2014. *Pew Research Center*, September 2014.

[12] M. N. Elliott, P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83, 2009.

[13] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):1955–1976, 2010.

[14] D. Irani, S. Webb, K. Li, and C. Pu. Large online social footprints–an emerging threat. In *International Conference on Computational Science and Engineering*, volume 3, pages 271–276. IEEE, 2009.

[15] S. Issenberg. *The victory lab: The secret science of winning campaigns*. Broadway Books, 2012.

[16] C. Jernigan and B. F. Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[17] H. Köpcke and E. Rahm. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69(2):197–210, 2010.

[18] T. Minkus, K. Liu, and K. W. Ross. Children seen but not heard: When parents compromise children's online privacy. In *Proceedings of the 24th International Conference on World Wide Web*. IW3C2, 2015.

[19] T. Minkus and K. W. Ross. I know what you're buying: Privacy breaches on ebay. In *Privacy Enhancing Technologies*, pages 164–183. Springer, 2014.

[20] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy, 2008*, pages 111–125. IEEE, 2008.

[21] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils. How unique and traceable are usernames? In *Privacy Enhancing Technologies*, pages 1–17. Springer, 2011.

[22] T. Pontes, G. Magno, M. Vasconcelos, A. Gupta, J. Almeida, P. Kumaraguru, and V. Almeida. Beware of what you share: Inferring home location in social networks. In *IEEE 12th International Conference on Data Mining Workshops*, pages 571–578. IEEE, 2012.

[23] T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, and V. Almeida. We know where you live: privacy characterization of foursquare behavior. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 898–905. ACM, 2012.

[24] I. Rubinstein. Voter privacy in the age of big data. *Wisconsin Law Review*, 2014.

[25] S. Sengupta. Update urged on children's online privacy. *New York Times*, September 15, 2011.

[26] S. Stecklow. On the web, children face intensive tracking. *Wall Street Journal*, September 17, 2010.

[27] C. Tang, K. Ross, N. Saxena, and R. Chen. What's in a name: A study of names, gender inference, and gender behavior in facebook. In *Database Systems for Adanced Applications*, pages 344–356. Springer, 2011.

[28] J. Turow, M. X. D. Carpini, and N. Draper. Americans roundly reject tailored political advertising at a time when political campaigns are embracing it. 2012.