RESEARCH REPORT SERIES
(Statistics #2012-13)

Exploring Re-identification Risks in Public Domains

Aditi Ramachandran[1]
Lisa Singh[1]
Edward Porter
Frank Nagle[2]

[1] Georgetown University
[2] Harvard University

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: September 21, 2012

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Exploring re-identification risks in public domains

Aditi Ramachandran
Georgetown University
ar372@georgetown.edu

Lisa Singh
Georgetown University
singh@cs.georgetown.edu

Edward Porter
US Census Bureau
edward.h.porter@census.gov

Frank Nagle
Harvard University
fnagle@hbs.edu

***Abstract*—While re-identification of sensitive data has been studied extensively, with the emergence of online social networks and the popularity of digital communications, the ability to use public data for re-identification has increased. This work begins by presenting two different cases studies for sensitive data re-identification. We conclude that targeted re-identification using traditional variables is not only possible, but fairly straightforward given the large amount of public data available. However, our first case study also indicates that large-scale re-identification is less likely. We then consider methods for agencies such as the Census Bureau to identify variables that cause individuals to be vulnerable without testing all combinations of variables. We show the effectiveness of different strategies on a Census Bureau data set and on a synthetic data set.**

I. Introduction

With the emergence of online social networks and social media sites, the increase in Web 2.0 content, and the popularity of digital communication, more and more public information about individuals is available on the Internet. While much of this information is not sensitive, it is not uncommon for users to publish some sensitive information, including birth dates and addresses on social networking sites. When potential adversaries have access to this large corpus of publicly available and potentially sensitive data, abuse can take place, leading to consequences such as fraud, stalking, and identity theft.

Cynics may question the value of protecting sensitive information that users are readily making public. Even so, a need to protect potentially sensitive data still exists for government entities and corporations. When agencies, such as the Census Bureau, release survey data, they need to be confident that the data cannot be used to re-identify survey participants. Not only are some data fields sensitive, e.g. income, but fewer individuals will participate in surveys truthfully if they are not confident that their identities will be protected.

This paper begins by presenting different strategies for re-identification given the large amount of public data available today. To better understand the level of difficulty associated with linking public information to other public data or to anonymized public information, we conducted two case studies, one involving Census data and one involving social networking data. The goals of each case study are slightly different. In the first case study, we are interested in determining how straightforward it is to link public data released from government agencies like the Census Bureau with public data that can be purchased from wholesale data sellers. In the second case study, we are interested in determining how straightforward it is to link two profiles on two different social networking sites, Twitter and Facebook, to create a more accurate profile of the individual. Based on the results of these case studies and others conducted in the literature, we conclude that targeted re-identification is not only possible, but relatively straightforward given the large amounts of publicly available data.

It would be beneficial for agencies, such as the Census Bureau to have a suite of tools to help them find these vulnerable individuals and distinct combinations of attribute features. Unfortunately, for large data sets containing a large number of attributes and a large number of records, exhaustively searching for individuals that are targets for re-identification is very costly. We present different heuristics that attempt to accurately identify attribute feature combinations causing individuals to be vulnerable without exhaustively searching all combinations and show their effectiveness on a Census Bureau data set and on synthetic data.

The contributions of this paper are as follows: (1) we present two different re-identification strategies using real world data sets; (2) we compare strategies for identifying variables that are causing individuals to be reidentified in synthetic data; (3) we analyze these strategies on real world data and conclude that the effectiveness of the strategies is very dependent on the type of data vulnerability present.

The remainder of this paper is organized as follows. Section II presents relevant literature. Section III presents a re-identification case study using public Census Bureau data, while section IV presents one using Twitter and Facebook data. In section V, strategies for finding variables that cause the vulnerability without testing all variable combinations are explored. Conclusions and future directions are presented in section VI.

II. Related Literature

A current area of research that often applies to crimes such as identity theft and fraud involves re-identification, or the process by which anonymized personal data is matched with its true owner. Even though potentially sensitive data is typically anonymized, re-identification approaches can sometimes be used to discover the identity of certain people in a data set [2], [1], [13]. Sweeney used a purchased voter registration list for Cambridge, Massachusetts and a publicly available medical data set to re-identify 87% of the people from the voter list using the attributes gender, zip code, and birth date [13]. Acquisiti and Gross demonstrated that using fields such as birth date, hometown, current residence, and phone number in conjunction can allow easier re-identification of a user and estimation of his or her social security number [1]. Narayana and Shmatikov [11] used an anonymized Twitter network and an un-anonymized Flickr network to re-identify nodes based on the similarities in graph structure. They were able to get a success rate of 30%. There has also been a great deal that has been written about re-identification risk. Hay et al characterized the risk of certain attacks based on the structural knowledge of a network dataset and demonstrated an approach to anonymizing network data that models aggregate network structure [6]. A recent emphasis on re-identification within health data has also yielded a study in which Dankar and El Emam develop and assess a re-identification risk metric for an adversary trying to re-identify as many records as possible. They were able to evaluate the metric using public and health datasets and demonstrated the growing need to assess the likelihood of large-scale patient re-identification in a disclosed health dataset [3]. While the cases studies presented in this paper are similar in spirit to some of this prior literature, both use different data sets.

Record linkage or record matching is another relevant area of research that attempts to map records in the same or different data sets to the same real world entity [12], [4], [14], [5], [7]. Record matching usually heavily relies on various string matching techniques and various distance metrics for determining the closeness of different attribute values. Traditional applications for record linkage, include duplicate record detection and medical record linkage. While we have leveraged some of the more basic string-matching techniques from the literature (see [14] and [5] for overviews), our studies differ from previous works since our focus is not just on the attack using record-linkage techniques, but also on exploring strategies for identifying variables that make individuals more vulnerable. Privacy-preserving record linkage deals with the process of maintaining individual privacy within databases that have been integrated from multiple sources and are shared across organizations [8]. While relevant, this work is complementary to the work presented in this paper.

III. Census Bureau Case Study

In this study, we attempted to link the American Community Survey (ACS) Public Use Microdata Sample (PUMS) that has been released by the Census Bureau with public data. Economists, psychologists, and sociologists use the PUMS for regression analysis and modeling applications to understand population demographics, changing social conditions, etc. It is not unusual for some released variables to be considered sensitive, e.g. income. By themselves, sensitive variables cannot be used to re-identify an individual, but when they are combined with an identifying attribute, sensitive information is revealed about the individual.

The remainder of this section discusses the privacy model used, the re-identification methodology, and the findings of the re-identification case study.

Algorithm 1 Re-Identification Algorithm - linkRecords

```
Input: $D', P, M$, $\alpha$, $\theta(D')$, $\theta(P)$
Output: $V$

for all $M_i$ in $M$ do
    compute $\delta(M_i)$
sort($\delta$)
N = $get\_combos(\delta, \alpha, M)$
for all $set$ in N do
    $att\_val\_combos$ = $get\_att\_val\_combos(set)$
    for all $combo$ in $att\_val\_combos$ do
        $count$ = $determine\_value\_count(combo)$
        if $count \leq \theta(D')$: then
            $match$ = $count\_records\_in\_P(combo)$
            if $match \leq \theta(P)$ then
                if $checkMatch()$ then
                    V.$add(combo)$
return $V$
```

A. Privacy Model

Given a data set $D(A_1, A_2, \ldots, A_m)$ containing m different attributes[1] and n records, assume that there are two types of attributes, identifying, A^I, and releasable, A^R. Identifying attributes cannot be disclosed since they map released attributes to a specific individual. Examples of identifying attributes include social security number and full name. Releasable attributes are those that by themselves cannot be used to identify an individual because multiple records contain instances of these attribute values. Examples of releasable attributes include gender and zip code.

An anonymized, sanitized version of D contains only the releasable subset of the original m attributes. We denote the anonymized version of D as $D'(A_1^R, A_2^R, \ldots, A_{nbr\_released}^R)$, where $nbr\_released < m$ and the number of identifying attributes, $nbr\_identify \geq 1$. A *re-identification privacy breach* occurs when one or more attribute values in A^I are determined for a specific tuple in D'.

To aid with the re-identification, we use a public data set $P(B_1, B_2, \ldots, B_r)$ that contains some releasable attributes present in D' and some identifying attributes present in D. Let $M(C_1, C_2, \ldots, C_s)$ represent the set of attributes that D' and P have in common. We will use record linkage methods to map records between D' and P using M to attempt to discover values for one or more attributes in A^I.

B. Re-Identification Methodology

At a high level, in order to find the identity of an individual in D', an attacker needs to find matching records in D' and P, where the number of possible matching records is small. Ideally, each record in D' would match only a single record in P. In practice, since D' does not contain identifying attributes, a small number of records in D' may match a particular record

[1] We will use the terms attributes, features, and variables interchangeably.

Number of Matching Individuals	**Number of Attribute Combinations**
< 10	14,741
< 5	5,227
1	926

TABLE II
RE-IDENTIFICATION - NUMBER OF MATCHING INDIVIDUALS IN ACS

in P. Records in D' are considered *vulnerable* if less than k individuals in D' match a single individual in P, where k is a small constant. The objective of this case study is to identify the set of vulnerable individuals in D using data in D' and P.

Our re-identification algorithm is described in Algorithm 1. The inputs to the algorithm are the released database (D'), the public database (P), the set of common attributes in P and D' (M), and the maximum size of attribute combinations we want to consider (α). α is used to reduce the computation cost by limiting the size of attribute combinations considered when M is large and the number of records in D' and P are large. Two other parameters, $\theta(D')$ and $\theta(P)$ are matching thresholds that are used to constrain the search space for matches in D' and in P. These parameters are generally very small, indicating that records should be further analyzed when only a small number of records match a specific set of attribute combinations in both D' and P. The output of the algorithm is the set of vulnerable individuals (V) found.

The algorithm begins by determining the distinguishing power of each attribute in M: $\delta(M_i) = |M_i|/|M_i|_{max}$, where $|M_i|$ is the number of distinct values for attribute M_i (the size of its domain) and $|M_i|_{max}$ is the maximum number of distinct values of any attribute in M. This algorithm makes the assumption that attributes with higher distinguishing power values are more likely to be involved in matching subsets that occur less frequently. For example, suppose we have three attributes, a, b, and c with 10, 100, and 500 distinct attribute values, respectively. Then this algorithm considers attribute combinations of b and c prior to considering combinations involving a. While this may not always be a good assumption, in practice, it is a reasonable approximation when testing all attribute value combinations is too costly. Therefore, once the distinguishing power of each attribute is determined, only those with a high value are combined to find vulnerable individuals.

The $get\_combos()$ method finds $\mathbf{N}$, the sets of attribute value combinations of size less than or equal to α. For each of the sets of attribute combinations in $\mathbf{N}$, the actual attribute value combinations that exist in D' are determined in $get\_attr\_val\_combos()$. For each attribute value combination, the number of matching records is then determined. If the number is less than $\theta(D')$, then the number of matching records in P is calculated. If the number of matching records in P is less than $\theta(P)$, then the $check\_match()$ method is called to determine whether other variables in M match for the records in question. Using these additional attributes, if a mapping exists from an attribute value combination in D' to a single record in P, then the mapped individual is identified as vulnerable and is added to V.

C. Re-Identification Results

The released ACS data set (D') contains 62 demographic and housing related variables.[2] To reduce the risk of identity disclosure, the data have been sanitized by the Census Bureau prior to public release using multiple disclosure avoidance strategies. This study focuses on three counties in three different states, California, Florida, and Texas.[3] The counties selected were based on availability of the data and because they were suburbs with a less transient residential population than urban communities. The total number of individuals in the ACS data set for the counties of interest is over 2 million people. Our goal is to attempt to find the correct names for one or more of these individuals using the re-identification methodology described in the previous subsection. For our public dataset P, we purchased demographic data for 700,000 individuals in specific counties from Wholesale Lists.[4] Data fields provided included name, date of birth, address, ethnicity, gender, and income.

After identifying the matching attributes in the released ACS data and in the public data set, we used the algorithm presented as Algorithm 1 to find vulnerable individuals. We set $\alpha = 7$, $\theta(D') = 10$ and $\theta(P) = 5$. Any attribute value combinations that had fewer than $\theta(D')$ individuals in the ACS data were flagged. We then matched those individuals to the public wholesale data set and flagged combinations that had $\theta(P)$ or fewer matching individuals. We then attempted to match individuals on other common attributes ($check\_match()$). If the individual was a complete attribute match, we considered the individual a possible re-identification match and added the individual to V. Finally, we hand-check the matches in V to records in D to determine our re-identification accuracy. Our accuracy rate is the number of individuals we correctly identified divided by the number of individuals in V.

Table I presents some statistics about the sample sizes for each state, the size of V, and the re-identification accuracy for the subset of ACS data we used. The overall vulnerability is defined by the correct number of vulnerable individuals divided by the number of surveys for the given state. We see that the accuracy ranges from 5% to over 60%, depending on the state. In general, the overall vulnerability for this population was less than 0.005%. Because these percentages are so low, we conclude that large-scale re-identification is unlikely when using basic re-identification techniques.

As previously mentioned, the number of matching attributes we used for this data set was seven or less. The most interesting were state, PUMA area[5], gender and age combinations. Table II shows the number of state, PUMA area, gender, and age combinations that exist in this data set having less than 10 individuals, 5 individuals, and finally, one individual. In other words, there are 14,741 combinations of state, PUMA area, gender, and age that have fewer than 10 people with that

[2] Released data details: http://www.census.gov/acs/www/

[3] For privacy reasons, we omit the year or counties used in this analysis.

[4] This data set was not private and can be purchased by anyone.

[5] Instead of counties, the Census Bureau uses PUMA areas as the lowest level of detail for geographic region

States	Nbr of Surveys	Nbr of Vulnerable Individuals	Nbr of Correct Vulnerable Individuals	Accuracy	Overall Vulnerability
California	1,028,566	233	54	23%	0.00005
Florida	547,847	43	27	63%	0.00004
Texas	675,158	113	6	5%	0.00001

TABLE I
RE-IDENTIFICATION ACCURACY RESULTS

combination. There are 926 individuals that have a unique combination of state, PUMA area, gender, and age in this ACS subset. It is interesting to note that even though this is the case, we could only re-identify 87 individuals accurately. There are a number of possible explanations, including the quality of the public data set and the fact that the ACS data set is a sample of the population and therefore, when rematched to a larger public population, more individuals that are not in the sample are matched. In other words, even though an individual is unique in the ACS data set, the individual may not be unique in the public data set. For example, if we only consider the 926 unique records, we find that those 926 records in D' matched just over 11,000 records in P. If we focus on the subset of individuals that have fewer than k=5 matches, then 56 individuals in D' match 158 individuals in P.

D. Discussion

This case study shows the viability of our re-identification procedure. It also highlights the need for high quality public data. Our public data set had a large number of records, but it also contained missing and inconsistent data. It is likely that this impacted the final re-identification accuracy. Even more importantly, it is interesting to note that when rematching to public data, the number of matches on common fields is high. To reduce this number and improve the matching accuracy, more matching fields are needed in M. Current public data sets from wholesale companies do not contain a large number of fields, generally, less than 10 that match to publicly released data sets. Therefore, large-scale re-identification using these publicly released data is more limited.

IV. SOCIAL NETWORKING CASE STUDY

While the previous case study used released public information from large entities, this one focuses on public information that is a product of users associating themselves with multiple social networking websites. Although the amount of public data differs amongst the websites, users are often unaware that information can be gathered and put together from multiple sources in order to create a more complete profile of themselves. While we are aware that users often purposely link their social networking accounts, this may be less desireable when considering other specific social networking websites (such as LinkedIn) and this case study is just an example involving two particular sites. In this case study, we ask the following question: how likely is it that a user from one online social network, e.g. Twitter, can be matched with some accuracy to a user on another online social network, e.g. Facebook, using public profile information? Does the likelihood increase when considering *social/friendship* attributes in conjunction with traditional re-identification attributes?

A. Re-identification Model

Our re-identification process tries to combine information from different sources to correctly identify a person across online social networks. Figure 1 depicts this process. Starting with a single user in social network S_1, we want to find the user in S_2 that has the same identity as the user in S_1. While most previous studies focus on traditional attribute matching, our focus is on understanding whether incorporating edge structure or friendship information from the source and target networks into the record matching process will improve the likelihood of a re-identification match.

More formally, we represent a general social network database as a graph $G(V,E)$ containing user nodes $V = \{v_1,\ v_2,...,v_n\}$ and friendship/follower edges between users $E = \{e_{ij} = (v_i, v_j)|v_i, v_j \in V\}$. We also assume that each user, v_i, has a set of public attributes, $A = \{a_1, a_2, ..., a_p\}$, e.g. name, age, etc. Let G^S be the starting or source graph and let G^T be the target or mapping graph. Our goal is to find an accurate match for users in G^S to users in G^T using attributes in A and possibly, edges in G^S and G^T. A privacy breach occurs when user v_i^S is mapped to a single user v_i^T.

B. Matching Algorithms

A basic approach for matching is to take every node in G^S and compare it to every node in G^T. Then take the node with the highest matching score and consider that a match. Unfortunately, even for small sizes of G^S and G^T, such a comparison is costly.

Therefore, we consider matching algorithms that use blocking variables. Blocking variables, B, are attributes that are used to segment the data set into groups or blocks, prior to matching. For example, *last name* may be considered a blocking variable. This means that for any node v_i^S, the only nodes in G^T that are considered for matching have the same last name as v_i^S. This allows us to narrow the

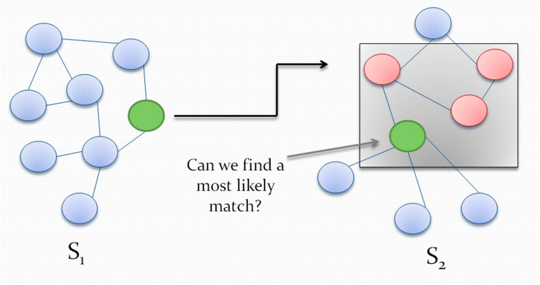


Fig. 1. Process of Re-identification Matching

Algorithm 2 FieldsMatch Algorithm

```
Input: $V^S$, $B$, $G^T$, $M$
Output: $S$

for all $v_i^S$ in $V^S$ do
    $V_B^T$ = Query target database for blocking matches
    for all $v_j^T$ in $V_B^T$ do
        scores[$j$] $\leftarrow$ 0
        for all field in $M$ do
            if $v_i^S$.field matches $v_j^T$.field then
                scores[$j$] $\leftarrow$ scores[$j$] + $weight_{field}$
    maxScore $\leftarrow$ max(scores)
    $S[i]$ $\leftarrow$ selectivity factor of maxScore
return $S$
```

Algorithm 3 FriendsMatch Algorithm

```
Input: $G^S$, $G^T$, $B$, $M$
Output: $S$

for all $v_i^S$ in $V^S$ do
    $V_B^T$ = Query target database for blocking matches
    scores[] $\leftarrow$ 0
    for all $v_j^T$ in $V_B^T$ do
        $match$ = num overlapping friends for $v_i^S$ and $v_j^T$
        scores[$j$] $\leftarrow$ $match$
    maxScore $\leftarrow$ max(scores)
    $S[i]$ $\leftarrow$ selectivity factor of maxScore
return $S$
```

potential list of users for a more detailed attribute matching. We consider two matching approaches: one based on user attributes in M (fieldsMatch) and one based on user friends in G(friendsMatch).

The metric used to evaluate each discovered match is called the *selectivity score*, where the selectivity score is defined as the inverse of the number of individuals with the highest matching score. This represents the probability that the algorithm can identify a given source user within the target data set. For example, if a particular matching result identifies three users in the target data set each with the same highest matching score, the selectivity score would be one-third. This gives a good indication of how likely it is that the given source user can be appropriately matched within the target domain. While other scoring methods can be considered, our focus is on a straightforward approach for evaluating matching success.

1) Node centric matching: The fieldsMatch algorithm matches users by comparing attributes in M on both G^S and G^T. Because users share similar information across social network sites, this algorithm focuses on matching that profile information. Algorithm 2 describes the approach. The input is the set of users V^S in the source graph G^S, the set of blocking attributes B, and the target graph G^T. The output is a list of the maximum selectivity scores for each user.

Using blocking attributes, the algorithm extracts a subset of the users V_B^T and considers these users to be potential matches. For each user in the set V_B^T, a matching score between this user and the given user from V^S is calculated based on the weighted matching attributes specified by the user. When the matching attribute of the starting user matches correctly to the corresponding field of the target user, the matching score between these two users is incremented by the weight associated with this attribute. After scores have been computed between the given user and each one of the target users in G^T, users with the highest score are selected. The selectivity factor corresponding to this score is computed and output.

2) Edge centric matching:: The friendsMatch algorithm, shown as Algorithm 3, is used to map users from G^S to G^T using the connectivity or friendship edge information that has been collected for each of these users. Given a specific user from G^S, using blocking attributes, we extract a subset of the users V_B^T and we consider these users to be potential matches. For each user in the set V_B^T, a raw count of how many overlapping friends exist between v_i^T and v_j^S. Friends of the users from each domain are counted as overlapping based on a simple name match. The user(s) in the set V_B^T that has the highest number of overlapping friends with the starting user is selected and the selectivity factor of this high count is computed and output.

C. Data Collection

In this study, G^S is created from a subpopulation of public Twitter users and G^T is created from public Facebook users. In order to acquire a sample of publicly available data, we created a custom data crawler to collect Twitter data. We began with a few random seeds and then used a breadth-first collection method. Our notion of friends on Twitter refers to the set of users a particular user has chosen to follow.

As the use for this data involves mapping our sample of Twitter users, v_i^S to Facebook users v_i^S, we searched for potential Facebook users through Google with the specified domain of Facebook.com. These public Facebook profile links were visited, where each link corresponds to one Facebook public profile and the user's profile information along with any friendship information was collected.

The breadth-first search on Twitter was started in late October 2009 and was running through early January 2010. The Facebook data was collected between January 2010 and April 2010.

Experiments were conducted on a set of 1600 Twitter users and these users were broken up into two distinct groups. One group of 500 users represent the Twitter users whose potential matches have more than eight friends. The second group of approximately 1100 users represent the group of Twitter users whose potential matches are based on eight friends.[6] Because the results we observed reflected the same trends for both test

[6] During the data collection phase, Facebook changed their public profile setup so that no more than a subset of a user's friends (at most eight friends) could be collected for a given user.

groups, we display more detailed results for the group with the more complete list of friends.

D. Experiments

Four variations of fieldsMatch were tested on both groups of Twitter users. The fields first name, last name, and location were chosen because the Twitter users in our study publicly shared this information and these attributes mapped directly to available fields of Facebook users with public profiles. The first experiment is simply a name match (called nameMatch), in which first name and last name act as blocking variables, and no matching attributes are specified. The second version of fieldsMatch, called fieldsMatch1, specifies an exact match of first name and an initial letter match of last name as blocking variables and an exact match of last name as a matching attribute. The third version, fieldsMatch3, has the same blocking set up as fieldsMatch2 and specifies a 'partial contains match' of last name as a matching attribute. The last version of fieldMatch uses an exact match of first name and a 'partial contains match' for last name as blocking variables and a 'partial contains match' for location as a matching attribute. The last test used in the experiments was friendsMatch, and the blocking variable setup was the same as fieldsMatch2.

For each of the five setups, three separate experiments were conducted. Statistics on average selectivity score and number of potential matches were collected first, for all the users in the group, second, for all the users who map to at least one potential match (blocking returns a set of at least one user), and third, for all users who map to at least two potential matches (blocking returns at least two users). Because it is difficult to assess the accuracy of the selectivity scores, or whether the Twitter user is indeed the same person selected as the match on Facebook, we hand validated the matches for a small set of users. The same five experiments were run on these users.

E. Results

Table III contains statistics for the groups of Twitter users that map to Facebook users with potentially more than eight friends. The displayed results in this table are for a threshold value of at least two potential matches. The range of selectivity scores for the four node-centric matching algorithms with at least two potential matches in Table III is .201 to .233 each having a standard deviation of around 0.17.

The friendsMatch algorithm has a selectivity score of 0.501, indicating that using connectivity information increases the likelihood of matching a user from Twitter to a user in the Facebook data set. Selectivity scores can be dependent on the number of potential matches, which for this experiment for all five algorithms range from 11.055 to 21.321.

Figure 2 shows the selectivity score for each individual Twitter user for each of the five matching algorithms. The graph shows the improved selectivity scores for the friendsMatch experiment. The noticeable increase in the number of users with a selectivity score of one shows that friendship information is a good attribute to use for re-identification matching across these data sets. Even though we did not hand validate the entire set of results, we hand validated some individuals who had a selectivity score of 1 when using the friendMatch algorithm. We found that in all cases, the friendMatch was accurate, while the nameMatch and fieldMatch2 were only correct in 20% of those cases.

Method	**Avg s_i for V^S with $t > 1$**	**stddev**	**Avg potential matches**	**stddev**
nameMatch	0.233	0.172	11.055	12.833
fieldsMatch1	0.231	0.174	21.321	38.601
fieldsMatch2	0.203	0.167	21.321	38.601
fieldsMatch3	0.201	0.166	18.441	39.076
friendsMatch	0.501	0.275	13.625	17.962

TABLE III
STATISTICS FOR TWITTER USERS MAPPING TO USERS HAVING MORE THAN EIGHT POTENTIAL FRIENDS

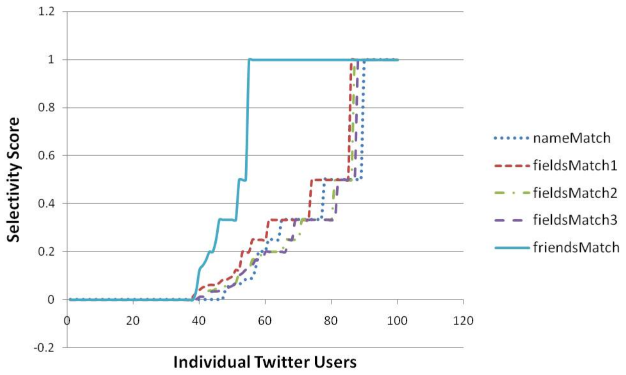


Fig. 2. Individual Selectivity Scores for Twitter Users Mapping to Facebook Users with More Than Eight Friends

F. Discussion

Despite the limitations on the data and the evaluation of the matching algorithms, the hand validated matches provide evidence that the edge-based attribute, friends, is a reasonable criterion on which to successfully match users across Twitter and Facebook data sets in comparison to regular user attributes such as last name and location. Although intuitively it seems as though location should improve selectivity scores, the lack of improvement can be attributed to the unstructured format for the data field on both Twitter and Facebook. Even with a partial match that makes use of state abbreviations and nicknames, the selectivity factor for users in the data set we used did not benefit from the location attribute.

V. IDENTIFYING VULNERABLE VARIABLES

Given the two case studies presented, it is clear that agencies and companies interested in releasing data to the public need to develop strategies for improving anonymization on large data sets. Our focus in this section is on *simple* methods for identifying variables that are vulnerable when the number of variables and records is too large to test the counts of all the variable combinations for vulnerability. We emphasize the term 'simple' as it is costly to consider a procedure that must scan the data numerous times for large data sets. Many methods have been proposed for anonymizing table data including k-anonymity, l-diversity, and t-closeness [13], [10],

[9]. We consider that research complimentary. Here we focus on identifying variables that cause vulnerabilities, not on the anonymization strategy once the variables are identified.

More formally, given a database $D(A_1, A_2, \ldots, A_m)$ with a set of attributes $A = \{A_1, A_2, \ldots, A_m\}$, our objective is to identify the set of *vulnerable attributes* $V \subset A$, where $|V| \ll |A|$, that contribute to making tuples in D vulnerable. A tuple is considered vulnerable if there are less than k other tuples in D with the same attribute values for attributes in A.

Based on the first case study, targeted attacks are more likely to be successful. Therefore, we focus on a more targeted attack, where an adversary knows a small number of attribute values for some specific individuals. The set of attribute values known to the adversary are referred to as the $CORE$ set of attributes, where $CORE \subset A$.

We consider two strategies in this section, core plus attribute(s) value count (CORE-AV) and core plus attribute(s) value threshold count (CORE-AT).

CORE-AV: Given a set of attributes known to the adversary ($CORE$), this method combines one or more attributes AV from $A - CORE$ with $CORE$, where AV is the set of attributes having the largest number of distinct attribute values and $|AV| \ll |A|$.

CORE-AT: Given a set of attributes known to the adversary ($CORE$), this method combines one or more attributes AT from $A - CORE$ with $CORE$, where AT is the set of attributes having the largest number of distinct attribute values with a tuple count less than min_count and $|AT| \ll |A|$.

We consider both CORE-AV and CORE-AT because CORE-AV chooses attributes with the largest domain, while CORE-AT looks for attributes with large domains that have values that appear less frequently. This takes into account the distribution of the attribute values rather than just the number of distinct values for a given attribute. Our approach is to determine the number of vulnerable individuals that can be found by considering the record values for the $CORE$ attributes and a small number of additional attributes. The number of additional attributes is pre-determined by the user and depends on the number of records and number of attributes in the data set. The reason this approach can be used to determine a good subset of vulnerable attributes is because of the following *superset vulnerable attribute* property.

Claim 1: Suppose an individual is vulnerable because of a vulnerable attribute A_v. The individual remains vulnerable when we consider attribute combinations for attributes $A_v and A_i$, where A_i is any non-vulnerable variable in A.
In other words, if a individual is vulnerable because he has a distinct birthdate in D, then he is still vulnerable when his birthdate is combined with a non-distinct attribute, e.g. gender.

A. Data sets

For this analysis, we attempt these strategies on a small synthetic dataset, as well as the ACS household data files mentioned in section III. The synthetic data set contains 10 attributes and 500 tuples. In the entire data set there are 5 vulnerable individuals. The vulnerabilities are based on a small number of attributes with the $CORE$. The ACS data used here is household data (approximately 1,040,000 records containing 75 attributes). For the $CORE$, we use very generic attribute values that an adversary targeting individuals would know. Specifically, for the household data, we use state, region, area, and type of home. Because we are using only a subset of the data, we are able to have a group of vulnerable individuals in the data. We say that an individual is vulnerable if the record for the individual is unique. Our goal is to find a set of variables that, when combined with the $CORE$ identify a large fraction of the entire set of vulnerable individuals in the data set.

B. Empirical results using synthetic data

Table IV shows the percentage of vulnerable individuals identified using our methods on the synthetic data set. Attribute rank is defined as the rank order of the particular attribute identified by the given method. For example, using CORE-AV, if attribute 3 had the largest domain and attribute 6 had the second largest domain, then attribute 3 would have attribute rank 1 and attribute 6 would have attribute rank 2. The method CORE does not utilize any additional attributes, so the attribute rank is N/A. While the percentages are very low for each individual method, if we combine the variables found across ranks 1, 2, and 3 for a method, i.e. the top three variables discovered using method CORE-AT with two attributes and the core, our approach determines 80% or 4 out of 5 of the vulnerable individuals and identifies three of the four variables used when creating the vulnerabilities. The three different combinations used with CORE-AT found distinct vulnerable individuals; therefore, while no one pair was able to find the majority of the vulnerabilities, using the method with multiple pairs successfully determined 80% of the vulnerable individuals.

C. Empirical results using released ACS Census data

In this data set, a single attribute by itself can not be used to identify vulnerable individuals. Figure 3(a) shows the percentage of vulnerable individuals that can be determined by each attribute in the data set. The x-axis represents the

Method	Nbr of Attributes with $CORE$	Attribute Rank	% Vulnerable Found
$CORE$	0	N/A	0%
CORE-AV	1	1	0%
CORE-AV	1	2	20%
CORE-AV	1	3	0%
CORE-AT	1	1	20%
CORE-AT	1	2	0%
CORE-AT	1	3	0%
CORE-AV	2	1, 2	20%
CORE-AV	2	1, 3	0%
CORE-AV	2	2, 3	20%
CORE-AT	2	1, 2	20%
CORE-AT	2	1, 3	40%
CORE-AT	2	2, 3	20%

TABLE IV
SYNTHETIC DATA VULNERABILITY IDENTIFICATION

Method	# of Attributes with $CORE$	Attribute Rank	% Vulnerable Found
$CORE$	0	N/A	0.06%
CORE-AV	1	1	52%
CORE-AV	1	2	39%
CORE-AV	1	3	45%
CORE-AV	1	4	18%
CORE-AV	1	5	1%
CORE-AT	1	1	39%
CORE-AT	1	2	52%
CORE-AT	1	3	45%
CORE-AT	1	4	18%
CORE-AT	1	5	3%
NAIVE	1	N/A	0.7%
NAIVE	1	N/A	0.3%
NAIVE	1	N/A	5%
NAIVE	1	N/A	0.01%
NAIVE	1	N/A	0.3%
CORE-AV	2	1, 2	61%
CORE-AV	2	1, 3	**90%**
CORE-AV	2	2, 3	79%
CORE-AT	2	1, 2	61%
CORE-AT	2	1, 3	79%
CORE-AT	2	2, 3	**90%**

TABLE V
CENSUS DATA VULNERABILITY IDENTIFICATION

attribute id, where each attribute is given a unique id between 1 and 75. The y-axis represents the percentage of vulnerable individuals found using only one of the attributes in the data set. Our figure highlights that only 9 attributes can be used to identify any vulnerable individuals. In the best case, an attribute can be used to find between 1 and 2 percent of the vulnerable individuals in the data set. Finally, using only the $CORE$ attributes, less than 0.1% of the vulnerable individuals in this data set can be determined.

Given this, we use our two proposed methods, CORE-AV and CORE-AT, to see if we can determine which attribute combinations lead to the identification of the majority of vulnerable individuals. Our first experiment compares three methods, CORE-AV using one additional attribute having the largest number of distinct values, CORE-AT using one additional attribute where $min\_count = 10$, and NAÏVE, a column randomly selected. Table V, compares the methods and shows the percentage of vulnerable individuals found for each method, where for each method we determine the top five variables containing attribute values causing vulnerabilities. The attributes for the NAÏVE method are chosen at random and do not have an attribute rank; therefore, the attribute rank for this method is shown as N/A. We see that both CORE-AV and CORE-AT outperform the random selection of attributes. This is not surprising given the large number of variables in this data set that are not involved in the vulnerability.

Both CORE-AV and CORE-AT find the same top four variables, but in different rank order. If we look at all the attributes with the CORE, we find that these four attributes are in fact the ones that lead to finding the largest number of vulnerable individuals. The fifth variable they find differs and is also not the actual fifth ranking attribute. Therefore, both methods determined the top 4 ranking attributes correctly and the fifth attribute incorrectly.

To better understand the distribution of vulnerable individuals when using the $CORE$ with a single attribute, Figure 3(b) shows the percentage of vulnerable individuals that can be determined by using the $CORE$ and every other attribute in the data set. We order the attributes from those leading to the identification of the largest number of vulnerable individuals to those leading to the identification of none of the vulnerable individuals. We see that the majority of variables are not good choices for re-identifying vulnerable individuals and that our simple methods do pick the best set of attributes in the CORE plus single attribute case.

Along with the $CORE$ plus single attribute case, we consider the $CORE$ plus two attribute case. We are interested in seeing the number of vulnerable individuals that can be determined using our methods with pairs of attributes and the $CORE$. Table V also shows these results. We can see that CORE-AT and CORE-AV are now able to identify the attribute combinations causing 90% of the vulnerable records. Figure 3(c) shows the percentage of vulnerable individuals found for all attribute pair combinations and the $CORE$ in this data set. This figure highlights the increase in vulnerable individual identification when two attributes are used with the $CORE$ when compared to using one attribute with the $CORE$.

By quickly identifying the variables that cause the vulnerabilities, agencies and companies can focus their effort on different anonymization strategies for those variables.

VI. CONCLUSIONS

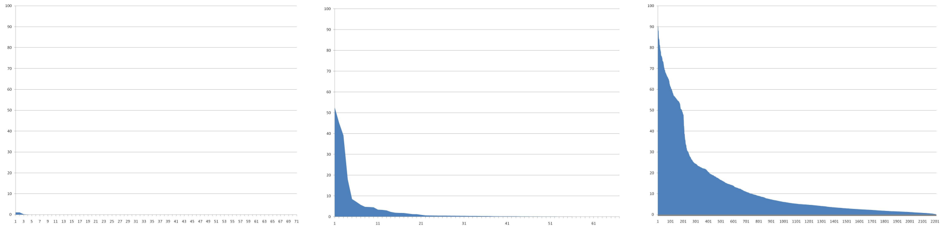
This paper presents two different cases studies for sensitive data re-identification, one involving one anonymized data file and a public data source and one involving two public data sources. We conclude that targeted re-identification using traditional variables and friendship variables is not only possible, but fairly straightforward given the large amount of public data available. However, our Census Bureau case study also indicates that large scale re-identification is less likely. We then consider methods for agencies such as the Census Bureau to identify variables that cause individuals to be vulnerable without testing all combinations of variables. We show the effectiveness of a simple method that determines variables that cause vulnerabilities using a very small number of attributes in a Census Bureau data set and a synthetic one. Future work will focus on testing our variable selection method on more data sets and determining vulnerable variables using both public anonymized data and social network data.

REFERENCES

[1] Alessandro Acquisiti and Ralph Gross. Information revelation and privacy in online social networks (the facebook case). In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society (WPES)*, November 2005.

[2] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, 2007.

[3] Fida Kamal Dankar and Khaled El Emam. A method for evaluating marketer re-identification risk. In *Proceedings of the 2010 EDBT/ICDT Workshops*, EDBT '10, 2010.



(a) Single Attribute (b) $CORE$ and 1 Attribute (c) $CORE$ and 2 Attributes

Fig. 3. Percentage of vulnerable individuals determined

[4] Pedro Domingos. Multi-relational record linkage. In *In Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining*, 2004.

[5] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. *IEEE Trans. on Knowl. and Data Eng.*, 19:1–16, January 2007.

[6] Michael Hay, Gerome Miklau, David Jensen, Don Towsley, and Philipp Weis. Resisting structural re-identification in anonymized social networks. *Proc. VLDB Endowment*, 1(1):102–114, 2008.

[7] Liang Jin, Chen Li, and Sharad Mehrotra. Efficient record linkage in large data sets. In *International Conference on Database Systems for Advanced Applications*, 2003.

[8] Alexandros Karakasidis and Vassilios S. Verykios. Secure blocking + secure matching = secure record linkage. *JCSE*, 5(3):223–235, 2011.

[9] Ninghui Li and Tiancheng Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *In Proc. of IEEE 23rd International Conference on Data Engineering ICDE'07*, 2007.

[10] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions Knowledge Discovery Data*, 1(1):3, 2007.

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

[12] Howard B. Newcombe and James M. Kennedy. Record linkage: making maximum use of the discriminating power of identifying information. *Commun. ACM*, 5:563–566, November 1962.

[13] Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[14] William E Winkler. Overview of record linkage and current research directions. Technical report, Bureau of the Census, 2006. Available at http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.