Identifying Individual Vulnerability Based on Public Data

John Ferro, Lisa Singh, and Micah Sherr
Georgetown University
Washington, DC 20057

***Abstract*—Companies and government agencies frequently own data sets containing personal information about clients, survey responders, or users of a product. Sometimes these organizations are required or wish to release anonymized versions of this information to the public. Prior to releasing these data, they use established privacy preservation methods such as binning, data perturbation, and data suppression to maintain the anonymity of clients, customers, or survey participants. However, existing work has shown that common privacy preserving measures fail when anonymized data are combined with data from online social networks, social media sites, and data aggregation sites.**

This paper introduces a methodology for determining the *vulnerability* of individuals in a pre-released data set to re-identification using public data. As part of this methodology, we propose novel metrics to quantify the amount of information that can be gained from combining pre-released data with publicly available online data. We then investigate how to utilize our metrics to identify individuals in the data set who may be particularly vulnerable to this form of data combination. We demonstrate the effectiveness of our methodology on a real world data set using public data from both social networking and data aggregation sites.

I. Introduction

Companies and government organizations frequently own data sets containing personal information about clients, survey responders, or users of a product. Sometimes these organizations are required or wish to release anonymized versions of this information to the public. Once a privately held data set is released, its privacy is protected only as long as unique individuals cannot be identified from among the released data. Even when explicit identifiers such as name or social security number are removed from data, re-identification can occur through the use of other unique sets of identifiers and record linkage [16].

An emerging concern regarding the release of even anonymized data is the proliferation of available public information sources that could be used to identify individuals. Although some social network data are protected through privacy settings, past work has shown that hidden attribute-values can be inferred using publicly available fields [10, 15]. In addition to data from social networks, the expeditious growth of the Internet has seen an incredible rise in the number of websites that specialize in providing and aggregating public information in an easily viewable and searchable format. These two phenomena have led to a significant increase in the availability of personal information.

This increase could have a large impact on the accepted methods of anonymous data publishing since individuals within a released data set will likely have an online presence. This presence will vary on a person by person basis, depending on the number of social networks a person has joined and the amount of information that is available about a person on the Internet. Currently, a significant concern in the field of anonymous data publishing is the linkage between publicly available web data and anonymous data being published.

To help address this concern, we propose methods to determine which individuals in a private data set are most vulnerable to linkage and other re-identification techniques due to the availability of public online information. We begin by introducing a methodology for assessing and evaluating how individuals' online presences can assist in the re-identification of those individuals in a released, anonymized data set. In particular, our techniques target individuals' *public profiles* – attribute-value pairs that are publicly available online and are useful for inferring private (non-public) values. We investigate how to determine the level of vulnerability of individuals when their public online information is combined with data an organization wants to publish.

More formally we state our problem as follows: Given a private data set D, identify those tuples whose vulnerability is higher than the expected vulnerability of tuples in D. We say an individual is *vulnerable* if the following three conditions hold: (1) a search for the individual across public websites returns one or more public profiles, (2) a small number of public profiles match on the attributes that are common between those public profiles and the individual, and (3) sensitive attributes about the individual can be discovered because of this match.

Given a set of individuals from a private data set D, with each individual having a set of attributes, an overview of our methodology is as follows:

- Search across public sites using an individual's attributes in order to find all of the possibly matching profiles for that individual;
- Rank all of the individuals based on their public profiles, using such information as how closely related an individual's public profiles match to that individual's original data and the number of public profiles found for an individual;
- Select those individuals with the highest rankings as the vulnerable set of individuals.

Through a detailed case study, we test this methodology. We use a purchased public data set and designate it as our "private data" that we plan to release. These data are then used to search across three social network and public information websites for the individuals within the data. We suppose that a certain subset of the individuals will be more vulnerable and have a larger online presence than other individuals. Our goal is twofold. First, we want to understand the online presence of a random individual. Second, we want to identify those individuals whose online presence is larger and more distinguishing than others in the private data set.

In summary, the contributions of this paper are as follows: (1) we present a methodology that can be used to identify a set of vulnerable individuals within a privately held data set; (2) we explore the factors that contribute to whether or not a specific individual is vulnerable; (3) we propose a simple hierarchical binning scheme that clusters individuals having similar vulnerability levels; and (4) we conduct a case study to analyze how the methodology works on real world data.

The remainder of this paper is organized as follows. Section II presents related literature. We present our methodology for discovering vulnerable individuals in Section III. Section IV presents our case study. We conclude in Section V.

II. Related Literature

In this section, we review the most relevant of the literature in the areas of record linkage, re-identification attacks, and data anonymization techniques.

Record linkage or record matching attempts to map records in the same or different data sets to the same real world entity [5, 9, 14]. Record matching generally 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. In this work, we have leveraged basic record linkage string matching techniques from the literature (see [6] and [17] for overviews).

The most relevant work to this paper is a study by Ramachandran et al. The authors demonstrate that it is possible to map released, sanitized private data to public data for a small subset of individuals in a data set [15]. The authors also consider individual vulnerability using public data prior to anonymizing the data for release. They focus on understanding the distinguishing power of an attribute in a private data set. In contrast, our goal is understanding the vulnerability of an individual based on public and private data matches.

A number of other re-identification studies using anonymized and un-anonymized data have been conducted [1, 2, 16]. Sweeney applies basic re-identification techniques to show that it is possible to link medical records and voter registration records in order to match names with private information such as diagnosis, procedures, and medications [16]. Acquisiti and Gross show that using basic demographic fields such as birth date, hometown, current residence, and phone number in conjunction can allow easier re-identification of a user and estimation of a social security number [1]. Narayana and Shmatikov [13] use an anonymized Twitter network and an un-anonymized Flickr network to re-identify nodes based on the similarities in graph structure. They are able to re-identify approximately 30 percent of the nodes in the graph based on the similarities in graph structure alone. A more recent study by Chaabane et al. [3] of 100,000 Facebook users finds that users are willing to share many attributes - 75% revealed gender, 57% revealed interests, and 23% revealed their current city.

Another thread of literature considers re-identification risk. Hay et al. characterize the risk of certain attacks based on the structural knowledge of a network data set [8]. Dankar and Emam develop and assess a re-identification risk metric for an adversary trying to re-identify as many records as possible in health data [4]. Liu and Terzi [11] calculate privacy scores for users who participate in different social networks. They consider the sensitivity of the field and the level of visibility of the disclosed information, where users in the network help determine the level of sensitivity of the field.

Finally, different approaches have been proposed for inferring private attributes from social network data. Zheleva and Getoor [18] use link-based classification to study the impact of friend attributes on the privacy of users. Using the attribute-values of friends in common groups, they infer a particular user's attribute-value. Chaabane et al. [3] use a Latent Dirichlet Allocation generative model to identify relationships between different interests specified by users. They show that Facebook users who are interested in similar topics with similar likelihoods have similar profile data. Lindamood et al. [10] use Facebook data and different Naïve Bayes classifiers to infer hidden political affiliation. Mislove et al. [12] use community detection metrics to infer attributes in two Facebook data sets. After identifying the community of the user, the authors determine the strength of the community using affinity and also consider the common attributes of the user community using modularity.

While all of this research is relevant to our problem, none of it directly investigates the problem we are exploring — that of mapping between individuals in a privately held, pre-anonymized data set and online public data for the purpose of identifying vulnerable individuals in the data set. This is an issue that must be faced when that privately held data set *needs* to be released for public use and the re-identification of individuals within the data is a concern.

III. Methodology

There are several steps that must be undertaken to identify the vulnerable individuals in a private data set. In this section, we propose an algorithm for determining individuals' vulnerability (Section III-A), introduce our metric for measuring the closeness of an individual to an online profile (Section III-B), describe our method of assessing an individual's overall vulnerability (Section III-C), and discuss ranking and binning strategies for classifying degrees of vulnerability (Section III-D).

Algorithm 1 Identify Vulnerable Individuals

```
1:  Input: $D$, $W$, $V_{\text{threshold}}$
2:  Output: $V$
3:
4:  for all $I_k$ in $D$ do
5:      for all $W_l$ in $W$ do
6:          $P^{I_k} = P^{I_k} \cup$ find_matching_profiles($I_k, W_l$)
7:      $\tau = \emptyset$
8:      for all $P_j$ in $P^{I_k}$ do
9:          $\tau = \tau \cup$ compute_data_match_score($I_k, P_j$)
10:     $S^{I_k}$ = compute_statistics($P^{I_k}, \tau$)
11: $R$ = determine_ranks($S$)
12: $V$ = select_vulnerable_individuals($R, V_{\text{threshold}}$)
13: return $V$
```

A. Vulnerable Individual Identification Approach

Given a private data set $D(A_1, A_2, \ldots, A_m)$ containing m attributes and n records, each tuple of m attribute-values represents an individual, I_k, in the data set, where $1 \leq k \leq n$. In addition, there is a set of l public websites $\{W_1, W_2, \ldots, W_l\}$ that can be chosen to search across. Besides the name of a site, information on what attributes can be gathered using the site is assumed to be known. Therefore, a website, W_j, is a set of attributes $(B_1, B_2, \ldots, B_h)$ that can be gathered from it. We denote the set of attributes at site W_j as $W_j(B_1, B_2, \ldots, B_h)$.

Our process for identifying a set of vulnerable individuals in a data set attempts to combine information from public websites with the information contained within the private data set. This identification process is described in Algorithm 1. The algorithm takes as input the privately held data set (D), a set of publicly available websites (W), and a vulnerability threshold ($V_{\text{threshold}}$) that indicates the level of vulnerability captured in the final vulnerable set. The output is the set of vulnerable individuals (V).

Our algorithm begins by searching for each individual I_k in the data set across all of the chosen sites (lines 4-6). The method find_matching_profiles() performs this search and takes two inputs: (1) the individual in D that is to be searched for, I_k, and (2) the site to be searched, W_l. This can be represented functionally as search : $I \times W \rightarrow P^*$, where $P^* = \{P_1, P_2, \ldots, P_t\}$, $t \geq 0$, and P^* is the set of public profiles returned for an individual. Specifically, find_matching_profiles() uses the search functionality of site W_l to find public profiles that match the attribute-values belonging to I_k. We denote the set of profiles returned by find_matching_profiles() for individual I_k as P^{I_k}.

Our use of websites' search functions is intended to reflect the behavior of an "adversary" who attempts to learn through online sources additional information about I_k that is not present in the released private data set. We remark that relying on a particular website's search functionality has two important effects. First, search queries may return multiple results, and hence it is possible (and as we show in Section IV, even likely) that $|P^{I_k}| > |W_l|$. Second, because the attribute-values in I_k may match multiple profiles on a website W_l, the set P^{I_k} may contain profiles that do not actually belong to I_k. (For example, a website may return several profiles based on the query "`name='John Smith'`", even though only one of those profiles may belong to the John Smith indicated by I_k.)

Lines 5-6 of our algorithm construct the set of public profiles P^{I_k} by searching across all sites in W. The algorithm then computes a *data match score* for each public profile in P^{I_k} (lines 7-9). The function compute_data_match_score() compares an individual I_k and a public profile $P_j \in P^{I_k}$ based on the values of their common attributes. We call the result of this comparison a public profile's data match score since it conveys how well the data found on the public profile matches with the data in the private data set. The specific composition of this data match score and how it is calculated is discussed later in this section. Conceptually, it represents the closeness of the online public profile to the individual I_k. Since I_k has a set of public profiles, once the data match score is calculated for each public profile, I_k will also have a set of data match scores, which we designate as τ.

After computing an individual's set of public profiles (τ), the method compute_statistics() calculates some summary statistics over the data match scores that comprise τ (line 10). As we discuss below, each summary statistic models some measure of the individual's vulnerability. Example statistics includes the median data match score, the number of scores (i.e., the number of public profiles returned), and the entropy of the scores.

Once the statistics are computed for all individuals in D, we then assign an overall vulnerability score to each individual $I_k \in D$ and *rank* the individuals according to their vulnerability (line 11). This ranking is performed by the method determine_ranks(), which first computes an ordering (ranking) of the individuals *for each summary statistic* returned by compute_statistics(). Then, the overall vulnerability score of an individual is computed as the sum of her statistic-specific rankings. For example, if the individual has a ranking of 3 for median data match score and 1 for entropy over data match scores, then her overall vulnerability score is 4. Finally, the individuals are ranked according to their overall vulnerability score, where R represents the final ordered list (rank) of individuals according to their vulnerability.

The method select_vulnerable_individuals() is then called with the set of rankings and the vulnerability threshold $V_{\text{threshold}}$ as inputs (line 12). This method groups the rankings together and then designates which groups of highest ranked individuals should be added to the set of vulnerable individuals, V. The tunable vulnerability threshold $V_{\text{threshold}}$ provides some control as to the level of vulnerability that is necessary to consider a particular $I_k \in D$ as being *vulnerable*.

In summary, this methodology ranks the vulnerability of individuals in a private data set. For each individual in the private data set, it identifies matching public profiles across different websites, computes a score (the data match score) that indicates how similar the attributes of a public profile

TABLE I
DATA MATCH SCORE EXAMPLE. ATTRIBUTE VALUE MATCHES ARE UNDERLINED.

Record	First Name	Last Name	Age	Gender	Data Match Score
Individual in D (ground truth)	Andrew	Smith	22	M	Not Applicable
Public Profile 1	Andrew	Jones	22	M	0.75
Public Profile 2	Amy	Smith	21	F	0.25
Public Profile 3	Andrew	Smith	22	M	1.00

are to those of the private individual, and then uses summary statistics of the data match scores to compute an overall vulnerability score for each individual. The vulnerability scores of individuals in the private data set are then clustered together based on similarity. Those with the highest vulnerability ranking are returned to the user.

B. Data Match Score

The data match score compares information that is gathered from a person's profile on a public website W_j with known information contained in the private data file D for an individual I_k. The goal in calculating a data match score is to generate a score that is representative of how closely related an online person is to the individual in the data set.

For any public profile P_i associated with individual I_k, let C_i be the attributes that are common to both the private data set D and the public profile P_i. We then define the data match score of P_i as follows:

$$\mathsf{data\_match\_score}(P_i) = \left[\sum_{c \in C_i} (\delta_c \times \alpha_c)\right] / \left[\sum_{c \in C_i} \delta_c\right]$$

where δ_c is an optional weight for attribute c, and α is a boolean match indicator that is 1 if the values of attribute c in I_k and P_i match (and is 0 otherwise). Note that the weights δ_c are used to specify the relative importance of attribute-value matches; for instance, we may consider a match of "first name" more revealing than "favorite band", since the former presumably changes far less frequently than the latter and is therefore more indicative of a match.

As an example, consider the sample individual and retrieved public profiles listed in Table I. In this example, a weight of 1 is used for all attributes. Also, for the sake of simplicity, all of the public profiles in this example have been collected from the same website, leading to all common attributes being the same across this set of public profiles. The first row of the table shows an individual who is in the private data set D. The remaining rows show the public profiles that have been found by searching online for this individual. Between the private data set and the public profiles that have been gathered, the attributes that are shared are First Name, Last Name, Age, and Gender. Since Public Profile 1 matches the individual in the data set for attributes First Name, Age, and Gender, its data match score is computed as $3/4 = 0.75$. (Matches are highlighted in the Table with underlining.) Public Profile 2, on the other hand, only matches the private data set for attribute Last Name, and so has a data match score of 0.25, while Public Profile 3 matches the individual in the data set across all common attributes and thus has a data match score of 1.

C. Overall Vulnerability

Lines 1-9 of Algorithm 1 produce a set of public profiles P^{I_k} and data match scores τ for each individual $I_k \in D$. In line 10, we next assign an *overall vulnerability* score to each individual to quantify her level of exposure if D is publicly released.

To compute the overall vulnerability score, we empirically evaluated different statistics over τ and P^{I_k} and found the following to be useful for assessing an individual's vulnerability:

- *the average, median, and maximum data match score* (intuitively, high data match scores indicate "successful" matches to online profiles);
- *the number of public profiles* (the number of returned public profiles is inversely proportional to vulnerability since an individual's true online profile may effectively "hide" in a large crowd of similar-looking profiles);
- *the standard deviation and Shannon entropy of data match scores* (both quantify the distribution of the data match scores across all returned profiles — low entropy or low variance among data match scores may indicate that no particular profile stands out from all profiles returned by the online search); and
- *the number of distinct fields that were collected across all public profiles* (this approximates the amount of additional information that can be learned about an individual through her online profiles).

We rank all of the individuals in the data set on each of these factors separately. This gives each individual seven different rankings. Finally, we calculate an individual's overall vulnerability score as the sum of the seven rankings. Weights can be added if certain statistics are considered more important than others; in practice, however, we find that summing the statistics yields reasonable estimates of vulnerability.

D. Ranking and Binning

The final step is to decide which of these ranked individuals should be added to the set of the most vulnerable individuals (line 12). A straightforward strategy is to label a particular percentage of individuals as vulnerable based on their vulnerability scores. The problem with this approach is that it gives no insight into the similarity in rankings of the individuals across the data set. Perhaps 2% of those in D are particularly vulnerable compared to the others in the data set. Or, perhaps that number is actually 40%. To avoid this problem, we propose using an approach that dynamically groups individuals having similar rankings.

The most common approach to grouping is binning data. A number of binning strategies exist (see Han et al. [7] for details), but the most widely used are equidepth and equiwidth binning. Equidepth binning places an equal number of individuals into each bin. While the bin size remains constant, the range of values within each bin may vary considerably. In contrast, equiwidth binning sets the range of each bin to be the same. This means that each bin may contain a varying number of individuals. While an improvement over equidepth binning, this binning strategy can result in some bins being empty and others being overly full. Given the weaknesses of these binning strategies, we propose a novel hierarchical binning strategy that does not predetermine the number of individuals in the bin or the width of the bin. Instead, it more intuitively clusters individuals with similar vulnerability scores. This hierarchical binning strategy can be viewed as a variant of hierarchical clustering [7] where the splitting criteria differs from the traditional clustering algorithm.

Our hierarchical binning strategy has two main steps: (1) building a vulnerability ranking tree using individual rankings and the standard deviation between these ranking; and (2) traversing the tree to get the most vulnerable set of individuals. We informally describe the algorithm using the example in Figure 1. The input to this algorithm is the set of rankings for all of the individuals (R) and the minimum standard deviation for splitting a bin ($V_{\text{threshold}}$). Intuitively, we want a bin to group together those individuals who have a similar vulnerability score. Standard deviation allows us to only keep individuals with a similar score in a single bin.

The algorithm begins with all the rankings being placed in the root node bin. The standard deviation is computed for this bin. In Figure 1 there are 10 vulnerable ranking scores placed in the root node. Because the standard deviation is above $V_{\text{threshold}}$, the bin is split. This process continues recursively until the standard deviation of each bin is below $V_{\text{threshold}}$. The leaf nodes of the tree represent the final bins. We consider the left-most leaf nodes to be most vulnerable. Our hierarchical binning strategy produces a vulnerability ranking tree that clearly identifies the most vulnerable individuals without arbitrary boundaries between groups and gives those releasing data further insight into the number of clusters and similarity of vulnerability scores for the entire data set. These strengths will be more evident when we compare the three binning strategies on real data in Section IV.

IV. Case Study

In this study, we show the viability and utility of our methodology. We acquire an offline data set containing demographic information and attempt to supplement that information using various online sources. As our "private" data set[1], we use a subset of the commercially available Wholesale Lists data that contains approximately 700,000 records of individuals' names, gender, ethnicity, income and home addresses (street, city, state, and zipcode, as well as their approximate latitudes and longitudes). From that data set, we construct our private data set D by randomly selecting 12,000 records from the Wholesale Lists data that belong to individuals in California, Texas, and Florida.

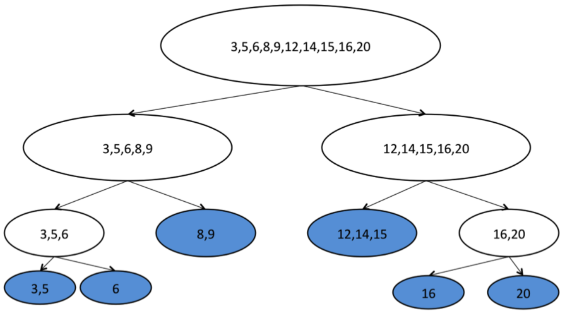

Fig. 1. Hierarchical binning tree.

A. Public Data Collection

To find public information about the individuals in our private data set, we query three online sources — LinkedIn, Whitepages, and Zillow. To better understand the variety of information that can be learned about an individual using partial knowledge, we purposefully chose diverse public data sources: LinkedIn is an online social network that specializes in professional networking. To match records from our private data set D with LinkedIn users, we use the site's search functionality and query for first and last name. To ensure that we retrieve only publicly accessible information, we use a fresh LinkedIn account that has no "contact" (friendship) relationships with the queried individuals. In contrast, Whitepages is a data aggregation service that specializes in contact information. We query Whitepages using individuals' names and states. Finally, Zillow is an online real estate marketplace that lists potentially sensitive information such as home property values. We match individuals to Zillow records by searching for their home addresses.

Characteristics of the public data sets. Since there is no explicit one-to-one mapping between individuals in the private data set and their online identities, searching the public websites for an individual I_k may produce multiple profiles $P^* = \{P_1, P_2, \ldots, P_t\}$. Figure 2 shows the cumulative distribution of the number of profiles returned from each website (as well as the combined total) for the individuals in the private data set. (We note that Whitepages limits the number of returned results to 100.) Unsurprisingly, the sites that allow more fine-grained queries based on some address

[1] In reality, the Wholesale Lists data set is commercially available and can be purchased by anyone.

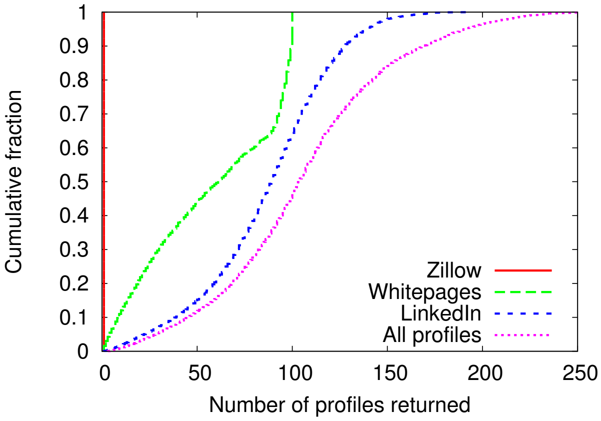

Fig. 2. Number of public profiles returned from searching for individuals on public websites.

TABLE II
COMMON ATTRIBUTES BETWEEN PRIVATE DATA SET AND ONLINE SITES.

Site	**Common attributes**
LinkedIn	first name, last name
Whitepages	first name, last name, street address, city, state, zip code
Zillow	street address, city, state, zip code, latitude, longitude

information (i.e., Zillow and Whitepages) return fewer profiles. In particular, Zillow always returned a single profile since the full address is used as the search criteria, and the median number of profiles returned by Whitepages is 61 (recall that only name and state are used to query Whitepages, resulting in more than one returned profile).

Each returned profile $P_i \in P^*$ contains attribute-values. Conceptually, since a public profile may contain information not present in D, the number of attributes present in an online profile serves as an indicator of the amount of additional information that might be revealed if D is released. To measure the amount of information exposed through online profiles, Figure 3 plots the cumulative distribution of the number of attributes for profiles returned from the three public sites, as well as the combined total ("All sites"). The cumulative distribution includes profiles collected for all individuals in D. Whitepages returns the greatest number of attributes and also has the greatest variance in the number of attributes returned. In contrast, LinkedIn and Zillow offer fewer attributes but are consistent in the number of attributes in each returned profile. Our results indicate that (1) a large number of additional attributes may be inferred from online profiles, and (2) the number of attributes returned varies both between sites, as well as amongst profiles within a site.

B. Data Match Scoring

The data match score describes the similarity between an online profile and an individual in D. To calculate data match scores for the online profiles, we use the common attributes between D and each site (see Table II). Note that since users may opt to not specify certain attribute-values in their online profiles, a site may return profiles that do not contain all possible common attributes; in these cases, only the common attributes that are present in both the profile and D are considered. We consider an attribute-value to match if the value from the private data is a substring or is equal to the corresponding attribute value in the online profile.

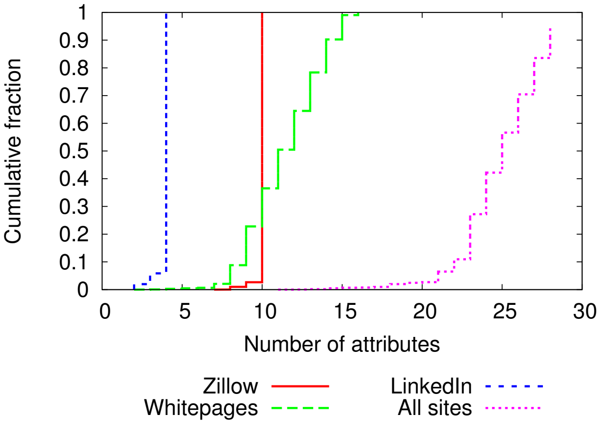

Fig. 3. Number of attributes for returned public profiles.

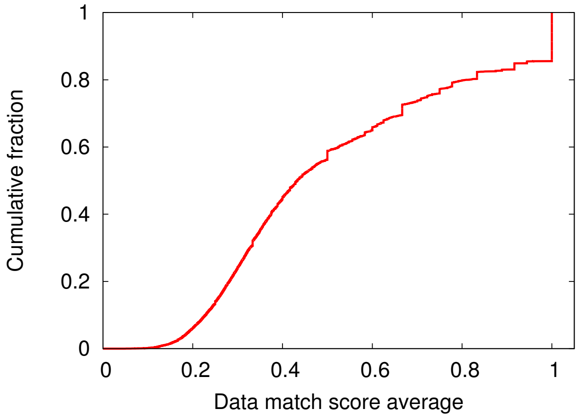

Fig. 4. Average of data match scores for all individuals.

For both LinkedIn and Zillow, approximately 90% of individuals have a maximum data match score of 1; roughly 80% of individuals have a maximum data match score of 1 on Whitepages. This indicates that a large fraction of individuals in D have at least one public profile on each of these sites that matches on all common attributes.

Figure 4 shows the cumulative distribution of average data match scores across public profiles from all three websites. Approximately 50% of individuals have an average data match score of less than 0.45. On the other hand, around 20% of individuals have an average data match score of at least 0.8; these correspond to individuals whose queries returned a small number of profiles and those profiles matched on almost every attribute. Overall, our results show that the returned profiles vary significantly in their matching accuracy. Given the variation, methods for quantifying vulnerability become particularly important for companies that need to release data.

TABLE III

STATISTICAL PROPERTIES OF DATA MATCH SCORES. THE "RANK ORDERING" COLUMN DESCRIBES THE SORT ORDER OF THE STATISTIC (GREATEST-TO-LEAST OR LEAST-TO-GREATEST) USED TO RANK INDIVIDUALS.

Statistic	Min/Max	Average	Variance	Median (1st/3rd quartiles)	Rank Ordering
Average data match score	[0, 1]	0.392	0.051	0.325 [0.238, 0.467]	Greatest to least
Median data match score	[0, 1]	0.375	0.117	0.333 [0, 0.5]	Greatest to least
Entropy of data match scores	[0, 119.212]	15.991	234.506	8.616 [1.431, 26.057]	Greatest to least
High data match score	[0, 1]	0.994	0.003	1 [1, 1]	Greatest to least
Standard deviation of data match scores	[0, 0.988]	0.328	0.025	0.317 [0.238, 0.422]	Greatest to least
Number of attributes	[14, 30]	24.799	5.565	25 [23, 27]	Greatest to least
Number of profiles	[3, 270]	71.201	2469.817	70 [25, 105]	Least to greatest

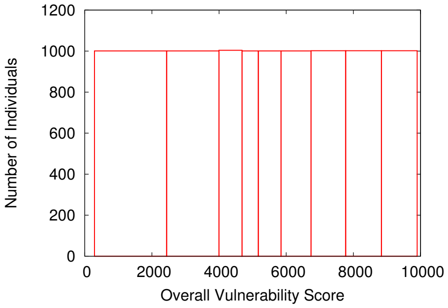

Fig. 5. Equidepth binning with bin size of 1000.

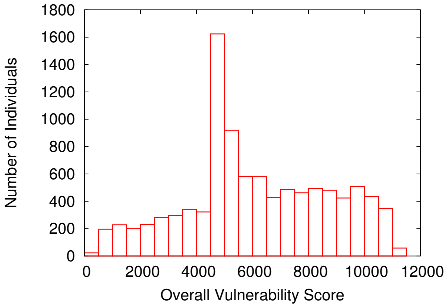

Fig. 6. Equiwidth binning with ranking range of 500.

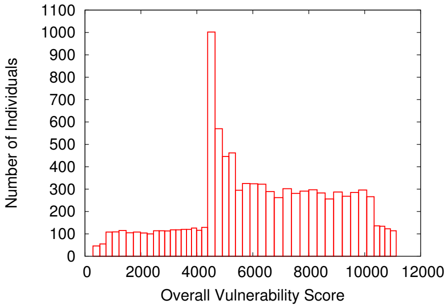

Fig. 7. Hierarchical binning with $V_{\text{threshold}}$ of 100.

C. Identifying Vulnerable Individuals

Using the data match scores, we rank the individuals in D according to their *vulnerability* — i.e., their ability to be identified in online public sources τ. To perform the ranking, we compute several statistics for each individual's data match scores. (Recall that an individual will have a data match score for each profile returned by the site.) Table III lists these statistics (leftmost column) along with each statistic's average, median, variance, and range. The rightmost column describes the sort order that is used when ranking individuals. (For example, individuals with high average data match scores are more vulnerable than individuals with low average data match scores, since high scores indicate close matches to I_k. In contrast, a large number of profiles indicates *less* vulnerability, since the true public profile belonging to I_k is hidden in a large set of returned profiles.)

To calculate an overall ranking for individuals, we first separately rank individuals on the basis of these statistics (e.g., average data match score, number of attributes, etc.). As an approximate indicator of an individual's vulnerability, we then sum these rankings for each individual, and rank order these sums to compute the final vulnerability ranking. Note that lower ranks reflect *more* vulnerable individuals. Conceptually, an individual's overall rankings indicates her level of exposure due to the public data, relative to other individuals in the private data set D. As discussed in Section III-D, a more customized rank order can be achieved by applying weights to each of the statistics listed in Table III.

Binning strategies. Before releasing a (potentially anonymized) data set, an investigator may wish to identify the *most* vulnerable group of individuals in the data set – i.e., those whom an adversary can learn the most amount of additional information by querying publicly accessible online sources. Although the overall vulnerability scores are sufficient to answer the question *who are the k most vulnerable individuals in D?*, it is also useful to group individuals in D according to their vulnerability. This is especially useful, for example, to identify those individuals in D that should be excluded from a public release of D.

We consider the equidepth, equiwidth, and hierarchical binning strategies described in Section III-D. Our goal is to assign individuals to bins (i.e., groups) such that an investigator can clearly distinguish between individuals that are vulnerable and those who are less exposed. For equidepth binning, we experiment with bin sizes containing 100, 500, and 1000 individuals. Figure 5 shows the result of this strategy using a bin size of 1000 individuals; similar results are achieved using bin sizes of 100 and 500 individuals, and are omitted for space. The y-axis gives the number of individuals present in the bin described by the x-axis. Equidepth binning allows analysts to easily identify certain groups (e.g., the group containing the first 1000 individuals). However, there is no guarantee that the individuals within this group have suitably similar rankings (i.e., overall vulnerability scores), or that the range of rankings covered by the bins is reasonably narrow.

For equiwidth binning, we use ranking ranges of 50, 100, 200, 500, and 1000; here, the ranking range denotes the maximum range in overall vulnerability scores in a particular bin. Binning with a ranking range of 500 produces the most useful results, and is depicted in Figure 6. While this strategy is useful for viewing and understanding the distribution of rankings within the data set, it does not take into account the similarities of rankings when performing the grouping. In particular, vulnerability scores may "cluster" around bin boundaries, making it difficult to obtain useful groupings of individuals.

The hierarchical binning strategy avoids these problems by actually considering the distribution of vulnerability scores within each bin. For hierarchical binning, we use standard deviation cutoffs of 50, 100, 200, and 500. We highlight the case of a cutoff of 100 in Figure 7. As explained in Section III-D, hierarchical binning creates a tree structure in which the root of the tree is a node containing all values that are to be binned. If the values in a node have a standard deviation less than the cutoff, the node is split, creating two new bins. The tree produced had 22 leaf nodes and 21 internal nodes. The final vulnerable set included 210 individuals with rankings from 240 to 980. The other binning strategies did not produce the same vulnerable set, instead producing a vulnerable set with either too few or too many individuals.

Overall, the vulnerability ranking tree that is created provides a useful overview of the vulnerability groupings in D. The leaf nodes give a micro-level clustering of the vulnerability rankings of individuals in the data set, while the internal nodes provide a macro-level view, highlighting trends of the individuals in the data set.

V. Conclusion

Owners of data sets containing personal information are sometimes required or wish to release anonymized versions of the data. This paper addresses the problem of identifying which individuals in the private data set are most vulnerable to re-identification using publicly accessible online sources.

Our techniques leverage a new metric called the *data match score* that quantifies the similarity between a record in the private data set and an online public profile obtained by searching a website. We describe how to combine data match scores from various websites to rank individuals in the private data set according to their expected level of information exposure. Finally, we present a binning strategy that groups individuals by their vulnerability, allowing the data owner to quickly discern which individuals are most at risk if the private data are to be published.

To evaluate the utility of our techniques, we conducted a case study using a commercially available database of demographic information (our "private data") and three diverse public online sources: a social networking site, a data aggregation service that specializes in contact information, and an online real estate marketplace. Our results indicate that (1) there is significant variation in the number of profiles returned when searching these sites for the individuals in the private data set, and (2) that the returned profiles also vary in the amount of information that they possess. By leveraging this variance, we can effectively *rank* individuals according to their level of exposure, allowing a data owner to quickly identify the most vulnerable individuals *before* the data are released.

Acknowledgements

We would like to thank the U.S. Census Bureau for access to the public wholesale data set used in this study. We would also like to thank Laura Zayatz and Edward Porter for their comments and feedback.

This work was supported in part by the National Science Foundation: CNS-1223825, CAREER CNS-1149832, CNS-1064986, and CNS-1204347 Any opinions, findings, conclusions, and recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the funding agencies.

References

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

[2] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *International Conference on World Wide Web (WWW)*, 2007.

[3] A. Chaabane, G. Acs, and M. Kaafar. You Are What You Like! Information Leakage Through Users' Interests. In *Network and Distributed System Security Symposium (NDSS)*, 2012.

[4] F. K. Dankar and K. El Emam. A method for evaluating marketer re-identification risk. In *EDBT/ICDT Workshops*, 2010.

[5] P. Domingos. Multi-relational record linkage. In *KDD Workshop on Multi-Relational Data Mining*, 2004.

[6] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19:1–16, January 2007.

[7] J. Han, M. Kamber, and J. Pei. *Data Mining, Second Edition: Concepts and Techniques*. Elsevier Science, 2006.

[8] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis. Resisting structural re-identification in anonymized social networks. *VLDB Endowment*, 1(1):102–114, 2008.

[9] L. Jin, C. Li, and S. Mehrotra. Efficient record linkage in large data sets. In *International Conference on Database Systems for Advanced Applications*, 2003.

[10] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *International Conference on World Wide Web (WWW)*, 2009.

[11] K. Liu and E. Terzi. A Framework for Computing the Privacy Scores of Users in Online Social Networks. In *IEEE International Conference on Data Mining (ICDM)*, 2009.

[12] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *ACM International Conference on Web Search and Data Mining*, 2010.

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

[14] H. B. Newcombe and J. M. Kennedy. Record linkage: making maximum use of the discriminating power of identifying information. *Communications of ACM*, 5:563–566, November 1962.

[15] A. Ramachandran, L. Singh, E. Porter, and F. Nagle. Exploring re-identification risks in public domains. In *Conference on Privacy, Security and Trust (PST)*, 2012.

[16] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.

[17] W. 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.

[18] E. Zheleva and L. Getoor. To Join or not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In *International Conference on World Wide Web (WWW)*, 2009.