Identifying Individual Vulnerability Based on Public Data - A Senior Thesis

A Senior Thesis
submitted to the Faculty of the
Computer Science Department
of Georgetown University

By John Ferro

Washington, DC
May 13, 2012

Senior Thesis

Department of Computer Science

Georgetown University

Identifying Individual Vulnerability Based on Public Data - A Senior Thesis

John Ferro,

Senior Thesis Advisors: Dr. Lisa Singh and Dr. Micah Sherr

Abstract

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. Previous research has shown that it is possible to uniquely identify an individual using only a small subset of these released attributes. Therefore, companies and government agencies use established privacy preservation methods, such as binning, data perturbation, and data suppression, prior to releasing data. However, in recent years, the amount of publicly available online information has grown at a staggering rate. A large amount of data on individuals can be found on various online social networks, social media sites, and data aggregation sites. This online data can be used in combination with released data to mitigate the privacy preserving measures and thus discover private information. Therefore, it is now imperative that companies and agencies consider these public data sources prior to releasing data.

This thesis introduces a methodology for determining how vulnerable individuals in a pre-released data set are 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 apply our metric to identify individuals in the data set which may be particularly vulnerable to this form of data combination. We demonstrate the effectiveness of our methodology on a data set containing 10,000 individuals using public data from three social network/public data sites. Our findings suggest that our approach will help companies or agencies identify vulnerable individuals and attributes in the data set on both the individual record level and the data set level.

Index words: Privacy, Privacy Preservation, Publishing, Social Network, Vulnerability

Chapter 1

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. In addition, it is important that released information is not used in order to infer additional information that can then be used for the re-identification of individuals. Even when explicit identifiers, such as name or social security number, are removed from the data, re-identification can occur through the use of unique sets of identifiers and record linkage [6]. An emerging concern is the increase in available public data. Although sometimes certain data from social networks is protected through privacy settings, past work has shown that some such data can be inferred using the publicly available fields [3]. 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 much more widespread and available public information in recent years.

This could have a large impact on the accepted methods of privacy preserving publishing. Traditional methods of protecting individuals' privacy in released data have been established on relational data [2, 4, 6]. With the rapid increase in the number of people using social networks, research on privacy preserving techniques that can be used on data gathered from social networks is emerging [6]. This past research has focused on how to protect users' privacy in social network data by providing access controls [1] and how established methods of privacy protection, such as k-anonymity, can be applied to publicly released social data [7]. Nowadays, it must be assumed that any attacker who is attempting to circumvent the privacy of released data has a readily available corpus of public data that can be utilized to supplement the attack. This data can range from traditional data fields providing information on an individual to the connections that an individual has made through his

or her social networks. An attacker can use this data in any form of record linkage or conventional re-identification attack.

Individuals within a released dataset 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 privacy preserving publishing is the composition of these web presences within the data that are being published. An analysis of individuals' online existences can show weaknesses in the traditional methods of privacy protection, as well as information on how to better protect the privacy of the data, by suppressing certain individuals or certain fields, for example.

This thesis explores how individuals' online presences can be assessed and evaluated based on how their presences would assist in the re-identification of the individuals. We investigate whether or not certain individuals can be found to be particularly vulnerable when their public online information is examined.

An overview of the methodology that we propose to identify the set of vulnerable individuals is illustrated in Figure 1. Given a set of individuals, with each individual having a set of attributes:

- 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.

In order to test this methodology, we take a purchased public data set and designate it as our "private data" that we are looking to release. This data is 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 a file.

The contributions of this thesis 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; and (3) we analyze how the methodology works on real world data.

The remainder of this thesis is organized as follows. Chapter 2 will explore literature that is related to this thesis. Chapter 3 will discuss the methodology that we use to discover vulnerable individuals. Chapter 4 will introduce the experiment that we perform using our methodology and analyze the results of this experiment. Conclusions will be presented in Chapter 5.

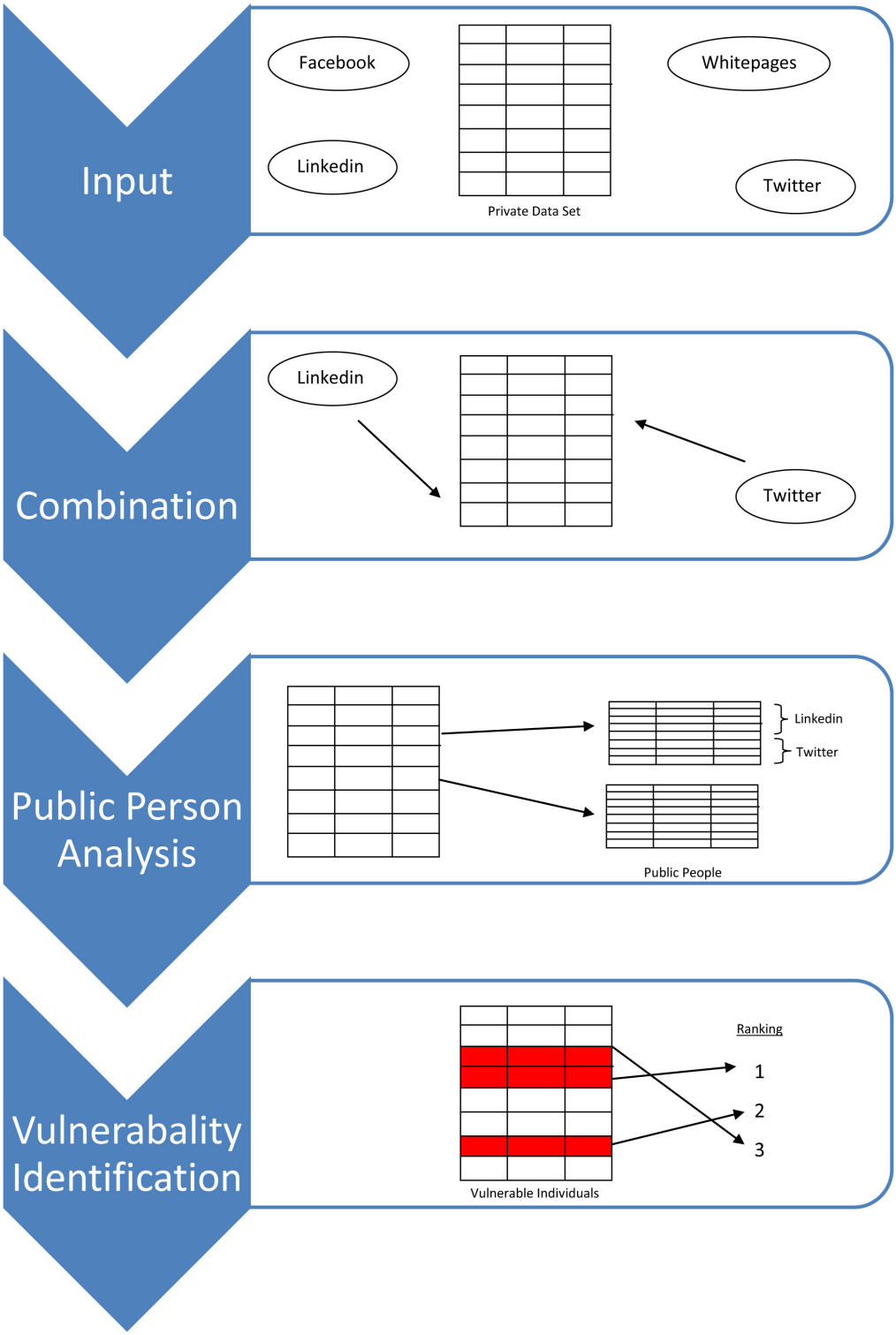

Figure 1.1: Vulnerable individual identification methodology.

Chapter 2

Related Literature

The fields of privacy preserving publishing and social network privacy contain a large amount of previous related works. Topics that need to be considered are re-identification attacks, data anonymization techniques, inferring private data on social networks, social network privacy preservation, and social network privacy controls.

2.1 Re-identification Attacks Using Public Data

Ramachandran et al. demonstrated that it is possible to map released, sanitized private data to public data [5]. They also demonstrate that it is possible to perform such a re-identification task using attribute matching as well as using a person's friends or connections within the social networks. The results also show that mapping individuals between social networks is possible with more certainty when the person's friends are used to help in the re-identification process.

2.2 Re-identification Attacks Using Quasi-identifiers

Sweeney examined both an attack on published anonymized data and also a privacy preserving method that defends against such an attack [6]. Sweeney showed that it was possible to link between medical records and voter registration records in order to match names with private information such as diagnosis, procedures, and medications. This type of re-identification makes use of attributes that are $quasi - identifiers$. Quasi-identifiers are attributes in one or more data sets that include both explicit identifiers such as a person's name, as well as attributes that can be used in combination with other attributes to uniquely identify individuals. In order to protect against re-identification attacks, Sweeney introduces the notion of k-anonymity. The k-anonymity requirement is a constraint placed on released data that helps to protect the amount of disclosure that can occur from quasi-identifying attributes. This k-anonymity requirement is met if and only if each sequence of quasi-identifying

values appears with at least k occurrences in the data set. This gives individuals in the data an anonymity of at least k individuals [6].

2.3 Social Network Privacy Preservation

Additional protection for published data may be required when the data comes from social networks. This is due to the fact that a host of new problems arise from social network data compared to relational data. For example, data relating to an individual's connections on a social network can be a uniquely identifying feature for an individual, but it is difficult to anonymize such data for release without tampering with the validity of the data. Zhou, Pei, and Luk survey this issue and claim that in order to effectively develop privacy preserving techniques for social network data, one needs to model the privacy information which may be under attack, the background knowledge that an adversary may use to attack the privacy of target individuals, and the usage of the published social network data so that an anonymization method can try to retain the utility of the data as much as possible while the privacy information is fully preserved [7]. They examine two categories of state-of-the-art anonymization methods on social network data, clustering-based approaches and graph modification approaches.

2.4 Methods for Inferring Private Information

It is important to look at previous work on general social network privacy protections in order to get a sense of the amount and quality of data that a potential attacker could gather. Fang and LeFevre have proposed an automated system for setting privacy settings which they call a Privacy Wizard [1]. They discuss how such a system is necessary due to flaws with the current privacy protections and settings among social networks. One particular example that they discuss is the Facebook "Privacy Settings" page where a user must manually assign friends to certain privacy settings. Fang and LeFevre's Privacy Wizard would take in a user's privacy preferences based on user input about privacy levels for certain friends. The Privacy Wizard would then take features from those already labeled friends and compare them to features extracted from unlabeled friends. Unlabeled friends would then be classified based on feature comparison, and the wizard would then be able to generate privacy settings for all of a user's friends based on these classifications [1].

2.5 Controls for User Privacy

Well placed privacy protections for social networks are extremely important since it is possible to infer private traits on social networks through the use of non-private traits. Lindamood et al. showed that this can be done using data from Facebook [3]. A variety of Naïve Bayes classifiers were used to infer the hidden political affiliation of an individual. One classifier that was tested only used attributes gathered from an individual's profile, a second only used friendship links between individuals in the network, the third classifier used was a combination of the first two, and the fourth was a traditional Naïve Bayes classifier. Lindamood et al. experimented with 35,000 users collected from Facebook and found that their combination classifier performed the best overall [3].

2.6 Comparison with our Work

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 becomes a concern.

Chapter 3

Methodology

The overarching goal of this thesis is to take a data set containing individuals and identify a subset of individuals who are particularly vulnerable to being identified using publicly available data.

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

There are several steps that must be undertaken in order to identify these vulnerable individuals. Initially, data must be collected from public websites that may relate to the individuals in the private data set. Secondly, this collected data must be compared with the data in the original data set and an empirical measure of this comparison must be computed. Finally, analysis must be performed on such an empirical metric to identify those individuals who should be considered vulnerable.

The remainder of this chapter discusses the procedure that we use to identify vulnerable individuals, and describes the most important parts of this process in greater detail.

3.1 Vulnerable Individual Identification Procedure

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 k is a number between 1 and 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 that can be gathered from it $(B_1, B_2, \ldots, B_h)$ where h is the number of attributes that can be gathered from W_j. We denote the set of attribute at site W_j as $W_j(B_1, B_2, \ldots, B_h)$ and a particular attribute B_k at site W_j as $W_j.B_k$.

identifyVulnerableIndividuals()
input: D, W, $vulnerabillity\_threshold$
output: V

foreach I_k in D:
  foreach W_l in W:
    $P_{I_k} = P_{I_k} \cup search\_for\_matching\_profiles(I_k, W_l)$
  end
  $\tau = \emptyset$
  foreach P_j in P_{I_k}:
    $\tau = \tau \cup compute\_data\_match\_score(I_k, P_j)$
  end
  $S_{I_k} = compute\_statistics(P_{I_k}, \tau)$
end
$R = determine\_ranks(S)$
$V = select\_vulnerable\_individuals(R, vulnerability\_threshold)$
return V

Table 3.1: Vulnerable Individual Identification

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 Table 3.1. The inputs when using this methodology are the privately held data set (D) and a set of publicly available websites (W), and a vulnerability threshold ($vulnerability_threshold$) which is a certain threshold indicating the level of vulnerability wished to be captured in the final vulnerable set. The output is the set of vulnerable individuals V.

Our algorithm begins by searching for each individual in the data set across all of the chosen sites. The method $search\_for\_matching\_profiles()$ performs this search and requires two inputs, the individual in D that is to be searched for, I_k, and the site to be searched, W_l. This can be represented functionally as $Search : I \times W \rightarrow P^*$, where $P^* = \{P_1, P_2, \ldots, P_t\}$ and $t \geq 0$. This indicates that P^* is the set of public profiles returned for an individual. An individual is provided to this method so that an individual's attribute values can be used in the search. The site needs to be provided since the site's own search features will be used, and the attributes that these search features take in to perform this search is site dependent. Each public profile, P_j, in this set corresponds to a profile that was found on site W_i using attribute values taken from I_k. Therefore, each individual, I_k, will have his or her own set of public profiles, P_{I_k}, that have been found by searching for that individual's data. This set can be represented as $P_{I_k} = \bigcup_{i=1}^{l} Search(I_k, W_i)$.

Once an individual, I_k, in D has been searched for on all sites in W, that individual's entire set of public profiles, P_{I_k} has been collected. For each of these public profiles, the method

compute_data_match_score() is called. This method takes in an individual, I_k, and a public profile, P_j, and compares the two based on the values of their common attributes. We call the result of this comparison a public profile's $DataMatchScore$ since it conveys how well the data found on the profile of a 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 chapter. 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 an individual, I_k has all the data match scores computed for their full set of public profiles, the method *compute_statistics*() is called taking in an individual's data match scores, τ, and public profiles, P_{I_k}, as inputs. Statistics for I_k are computed based on these inputs. The set of public profiles needs to be passed into this method in addition to the data match score since information about these public profiles is calculated that was not part of the data match score calculation. This is because the data match score only compares an individual with one public profile. However, more "macro" information about the composition of public profiles found for an individual is needed to determine an individual's overall vulnerability. The method *compute_statistics*() returns these calculated statistics for an individual, I_k, and stores them as S_{I_k}. S_{I_k} is part of the set S where $S = \{S_{I_k} \mid 1 \leq k \leq n\}$.

Ranking can occur once the statistics are computed for all individuals in D. This ranking is performed by the method *determine_ranks*() and the set S is an input to this method so that it has access to the statistics for all individuals. First, a ranking is calculated for each statistic and then each of these initial rankings is summed for an individual in order to compute an overall ranking based on multiple statistics. This set of rankings, R, is returned by the method, and each individual in D has one ranking in R.

The method *select_vulnerable_individuals*() can then be called with the set of rankings and a vulnerability threshold as an input. This function groups the rankings together and then designates which groups of highest ranked individuals should be added to the set of vulnerable individuals, V. The vulnerability threshold is given as input to give a user of this methodology some control as to what level of vulnerability is used to add individuals into the final set of vulnerable individuals.

3.2 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, 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. Our data match score is shown in the following equation.

$$data\_match\_score = \left[\sum_{i=1}^{x}(\delta_i \times \alpha_i)\right] \div \left[\sum_{i=1}^{x}\delta_i\right]$$

Between an individual from the private data set, and a public profile gathered from the internet, there will be a set of x attributes that are common between the two. If a particular attribute value, for attribute B_i in this set of common attributes, is found to be a match between the individual in the data set and the public profile, then value of α_i is set to 1. However, if there is not a match, then α_i is set to 0. The value of δ_i is a weight that can optionally be assigned to attribute B_i in order to make certain attributes more or less important in this calculation. For example, the attribute "last name" may seem like it should be a more important indicator of a match, and so should be given a larger weight than the attribute "favorite color". These weights are determined a priori by the analyst making use of this methodology. The first step of calculating the data match score is to compute the summation of $(\alpha_1 \times \delta_1, \alpha_2 \times \delta_2, \ldots, \alpha_x \times \delta_x)$. Then, this summation is divided by the summation of $(\delta_1, \delta_2, \ldots, \delta_x)$, the weights for all common attributes.

Table 3.2 demonstrates how the data match score is calculated, using a small example. 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. In Table 3.2 the first row shows an individual who is in the private data set. The remaining rows contain information on 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. Therefore, in this example, x is 4. Since public profile 1 matches to the individual in the data set for First Name, Age, and Gender, public profile 1's data match score is computed as 3 / 4 = 0.75. Public profile 2, on the other hand, only matches 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.

Person ID	First Name	Last Name	Age	Gender	Data Match Score
Individual in D	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

Table 3.2: Data Match Score Example

3.3 Ranking and Binning

The purpose of ranking individuals is to order all of the individuals in the data set based on how vulnerable an individual is to being identified using publicly available online data. Therefore, in order to determine what should compose a ranking system, it must be determined what factors influence how vulnerable an individual is. Many of these factors can be expressed through statistical analysis of an individual's data match scores; however, other factors be must calculated separately using an individual's set of public profiles.

We empirically evaluated different statistics and found the following to be useful for assessing an individual's vulnerability: the average of the data match scores, the median of the data match scores, the entropy of the data match scores, the largest data match score, the standard deviation of the data match scores, the number of distinct fields that were collected across all public profiles, and the number of public profiles.

Each of these statistics is an important component of an individual's overall vulnerability.

- The average of the data match scores represents on average, how well any given public profile matches with the individual in the private data set. The higher this average score is, the more closely related all of the public profiles are to the individual.

- The median data match score indicates much the same vulnerability score, but is a more accurate average value for certain data match score distributions.

- The largest data match score represents how closely the most related public profile is to the individual.

- The entropy and standard deviation help to describe the distribution of data match scores. This distribution is important in determining an individual's vulnerability since an individual would

Individual	0	1	2	3	4	5	6	7	8	9
Ranking	1	1	2	2	3	4	5	6	7	8

Table 3.3: Binning Example Data Set

be more vulnerable if there are public profiles that stand out from the rest of the population of public profiles.

- The number of public profiles returned for a given individual, since more public people returned means that an individual has a more anonymous .
- The number of fields gathered from public profiles, since this number indicates how much additional information can be gathered on the individual by looking at online sources.

We rank all of the individuals in the data set on each of these factors separately. This gives each individual seven different rankings. An overall ranking is then computed by adding up all of these individual rankings. This overall ranking is indicative of how vulnerable an individual in the data file is. Weights can be added if certain statistics are considered to be more important.

The difficulty is then deciding which of these ranked individuals make it into the most vulnerable set of individuals. For this to be done, individuals need to be grouped together on the basis of their rankings, in order to get sets of closely ranked individuals. Then, the set that contains the highest ranked individuals will be the final most vulnerable set.

To perform this grouping, we propose using several different binning strategies. One such strategy is equidepth binning, which places an equal number of individuals into each bin, but the bins are of varying sizes. A second strategy is equiwidth binning. This sets an equal size to each bin, and then places the varying number of individuals into the correct bins. Given the weaknesses of the first two binning strategies, we propose a novel hierarchical binning strategy. With this strategy, a minimum standard deviation for splitting a bin is decided upon. Then with the entire set of rankings as a starting point, a recursive method is used which calculates the standard deviation for the current bin, and if it's above the minimum threshold, splits the bin in half, and then calls itself on each of these halves. Once the standard deviation of a bin falls below the threshold, the splitting stops.

Table 3.3 shows an example data set that can be grouped together by these binning strategies. For this example, a depth of 2 will be used for the equidepth strategy, a width of 2 will be used for the equiwidth strategy, and a standard deviation cutoff of 1 will be used for the hierarchical strategy.

Bin Number	Individuals	Rank Range
1	0-1	1
2	2-3	2
3	4-5	3-4
4	6-7	5-6
5	8-9	7-8

Table 3.4: Equidepth Binning Results

Bin Number	Individuals	Rank Range
1	0-3	1-2
2	4-5	3-4
3	6-7	5-6
4	8-9	7-8

Table 3.5: Equiwidth Binning Results

With equidepth binning, each bin that is created has two individuals placed in it. Table 3.4 shows that with this strategy, 5 bins are created. Since there are two individuals that have the ranks 1 and 2, bins 1 and 2 only have these values in them. The rest of the bins each have two different ranks in them. With equiwidth binning, each bin that is created has a range of ranks that are all placed in it. In this example the width of this range is 2, and so the first bin that is created has four individuals in it, since both ranks 1 and 2 have the two individuals that have those rankings. The rest of the bins each contain two differently ranked individuals, as seen in Table 3.5. The hierarchical strategy starts with one bin that contains all values within it, in this case, all indices 0-9. The standard deviation of this bin is then calculated, and comes out to 2.51. Since this is above the threshold of 1, the bin is split into two bins, individuals 0-4 and 5-9. The standard deviation of the first of these two bins is calculated to be 0.84. This is below the threshold and so this bin remains as is. However, the standard deviation of the other bin is 1.58 and therefore splits again into a further bin, of individuals 5-6 and 7-9. The bin with indices 5-6 has a standard deviation of 0.71 and the other bin has a standard deviation of 1, and so they both stop binning. Now that there is no more splitting occurring, the final bins have been reached. The contents of these final bins are shown in Table 3.6.

Bin Number	**Individuals**	**Rank Range**
1	0-4	1-3
2	5-6	4-5
3	7-9	6-8

Table 3.6: Hierarchical Binning Results

Chapter 4

Experiments and Results

4.1 Private Data Set

To imitate a private data set, we used a set of demographic data for 700,000 individuals that was purchased from a whole sale company that was provided by the Census Bureau. Attributes that were provided within this data set included name, street address, city, state, zip code, latitude, and longitude. We took a subset of 10,000 individuals from this data set to be used as our private data set in our experiments.

4.2 Public Data Collection

We chose to search three websites in order to find public information about the individuals in our data set. These websites were Whitepages, LinkedIn, and Zillow. Whitepages is a data aggregation site specializing in contact information. Linkedin is a social networking site focused on business and job networks. Zillow is data aggregation site containing information on addresses and properties. In order to search for individuals across these websites, two main methods were used. For Zillow and Whitepages, the sites' own search functions were used. For LinkedIn, a google search was performed with the specified domain of linkedin.com. In either case, performing a search returned a list of relevant links which were then explored in order to gather data about each public profile. Since a different search function was being used on each site, these searches also accepted different inputs on which to perform the searches. LinkedIn was searched based on first name and last name. Whitepages was searched based on first name, last name, state, and zip code. Zillow was searched based on street address, city, state, and zip code.

The number of public profiles returned varied based on the individual being searched for, as well as the website that was being searched. Figure 4.1 depicts the cumulative distribution function of the number of profiles returned, on a per site basis, but also for all of the public profiles returned for an individual. Zillow's searches only returned a single public profile, since the Zillow search

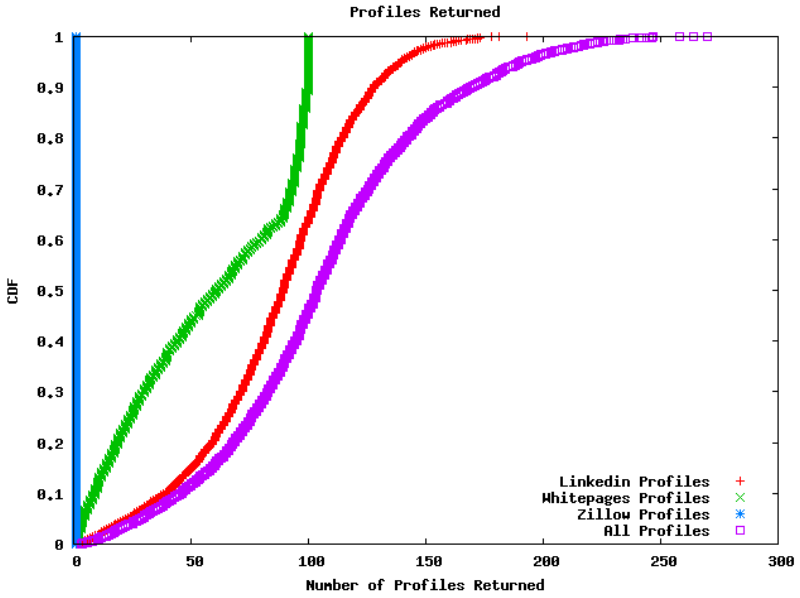

Figure 4.1: Number of Public Profiles Returned from Searching for Individuals on Public Websites

function mapped the single address being searched for to a single public profile. However, searches on both LinkedIn and Whitepages returned a varying number of profiles based on the individual being searched for. It is worth noting that due to the Whitepages search function, the maximum number of individuals that could be returned by a search on Whitepages was 100. The total number of public profiles returned for all individuals was around 650,000, and as can be seen in Figure 4.1 the median number of profiles returned was around 100 per individual.

The attributes that could be collected from public profiles was dependent on the site that the public profile came from. Figure 4.2 shows the cumulative distribution function of the number of attributes that were gathered for an individual for public profiles returned from each site, as well as for all public profiles returned for that individual. Whitepages had the most attributes that could be gathered, but also the most varying number of attributes that were found. A much more constant number of attributes were found on both LinkedIn and Zillow. For these two sites, the maximum number of attributes that could be gathered were gathered for 90% of all individuals. Examples of attributes gathered from Linkedin are first name, last name, and education. Examples of attributes gathered from LinkedIn are first name, last name, phone number, and age range. Examples of attributes gathered from Zillow are street address, city, home value, latitude, and longitude.

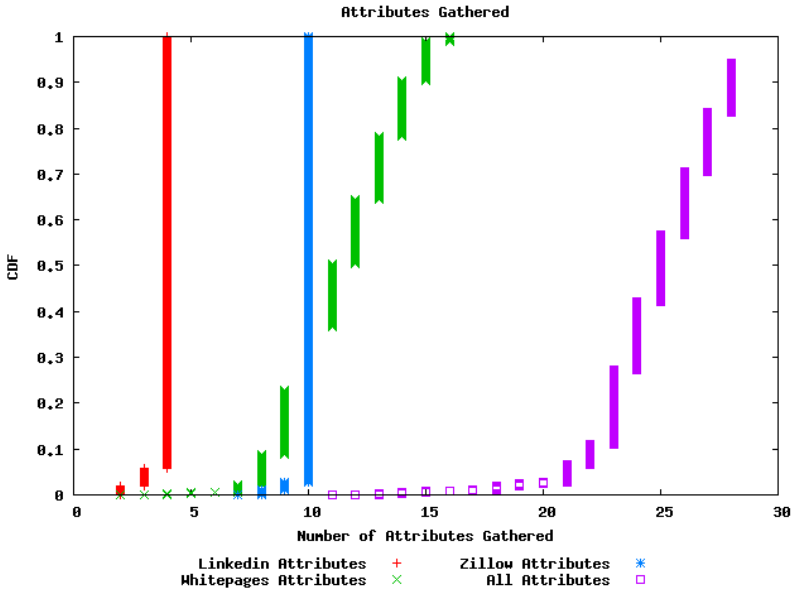

Figure 4.2: Number of Attributes Gathered from Public Profiles

4.3 Data Match Scoring

Once all public profiles were collected, we used the data match scoring described previously in order to give a score to each public profile. When calculating the data match score, the common attributes between the individual in the private data set and the public profile need to be known. With our data, the common attributes for profiles from Linkedin were first name and last name, the common attributes for profiles from Whitepages were first name, last name, street address, city, state, and zip code, and the common attributes for profiles from Zillow were street address, city, state, zip code, latitude, and longitude. However, these were the maximum possible common attributes for each site. For any give public profile, only a subset of these attributes may have been present, and so the data match score only uses these present common attributes. To compare the attribute value between the private data set and a public profile, we made use of the python regular expression search function. If the attribute value of the private data set was found within the attribute value of the public profile, this attribute was designated as a match. This was useful for negating the effects of extra data from excess html code mistakenly gathered during the data collection process, but does not catch such issues as spelling mistakes. A more advanced comparison function could be used in order to provide for correct matching despite such errors.

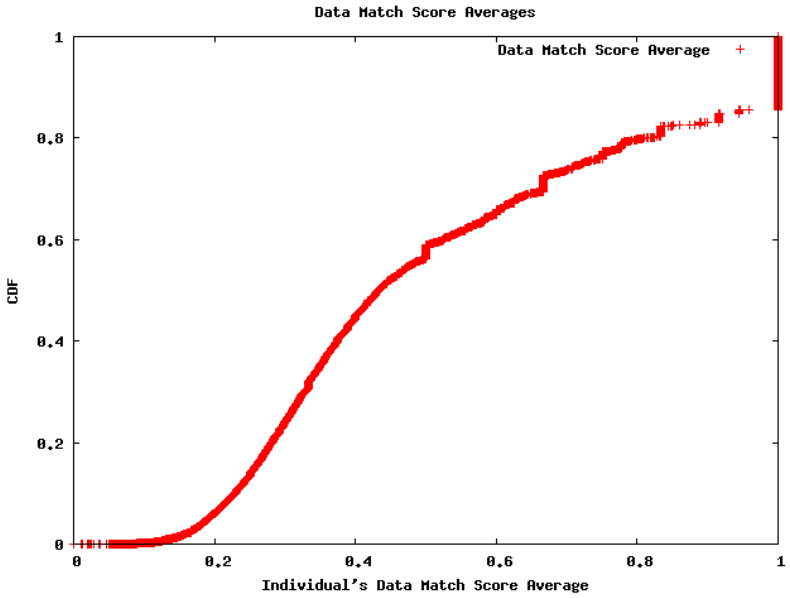

Figure 4.3: Average of Data Match Scores for all Individuals

For each individual's set of data match scores, the maximum data match score calculated for each was most often 1. From Linkedin and Zillow, approximately 90% of individuals had a maximum data match score of 1, and from Whitepages, approximately 80% of individuals had a maximum data match score 1. However, while this maximum data match score was similar for all individuals, the average of individual's data match scores showed much more variation. This variation is shown in Figure 4.3, a cumulative distribution function of all individuals average data match score. This average was calculated across public profiles from all three websites. As can be seen from this figure, around 50% of individuals had an average data match score of less than 0.45. On the other hand, around 20% of individuals had an average data match score of at least 0.8. These individuals would be those who had a small number of public profiles returned when they were searched for, but matched on almost every attribute.

4.4 Statistical Computations

After all of the data match scores were calculated, a number of statistics were computed for each individual that would then be used to rank the individuals. These statistics were the average of an individual's data match scores, the median of an individual's data match scores, the entropy of an individual's data match scores, an individual's highest data match score, the standard deviation of

Statistic	Min Value	Max Value	Range	Average	Variance	Median	First Quartile	Third Quartile
Average Data Match Score	0	1	1	0.392	0.051	0.325	0.238	0.467
Median Data Match Score	0	1	1	0.375	0.117	0.333	0	0.5
Entropy of Data Match Scores	0	119.212	119.212	15.991	234.506	8.616	1.431	26.057
High Data Match Score	0	1	1	0.994	0.003	1	1	1
Standard Deviation of Data Match Scores	0	0.988	0.988	0.328	0.025	0.317	0.238	0.422
Number of Attributes	14	30	16	24.799	5.565	25	23	27
Number of Profiles	3	270	267	71.201	2469.817	70	25	105

Table 4.1: Analysis of Statistics

an individual's data match scores, the number of attributes that were collected from public profiles for an individual, and the number of public profiles gathered for an individual. These were computed both for each site and across all sites. To get a sense of the distribution of these statistics across all individuals, the lowest value, highest value, range, average, variance, median, first quartile, and third quartile were calculated. Table 4.1 presents these calculations for when they were computed across all sites.

4.5 Identifying the Vulnerable Individuals

With these statistics computed, we used them to rank individuals. To calculate an overall ranking for individuals, we first separately ranked individuals on the basis of these statistics. Then we sum these rankings, in order to compute a final vulnerability ranking. At this point, we applied all three binning strategies in order to group individuals together based on their rankings. For equidepth binning, we looked at bin sizes of 100, 500, and 1000. Figures 4.4, 4.5, and 4.6 show these binnings. As can be seen, this binning strategy equally distributes the same number of individuals to each been, and so are of the same size, except in cases when there are ties. This form of binning groups individuals in such a way that easily enables picking, for example, the group containing the first 100 individuals. However, there is no guarantee that the individuals within this group have suitably

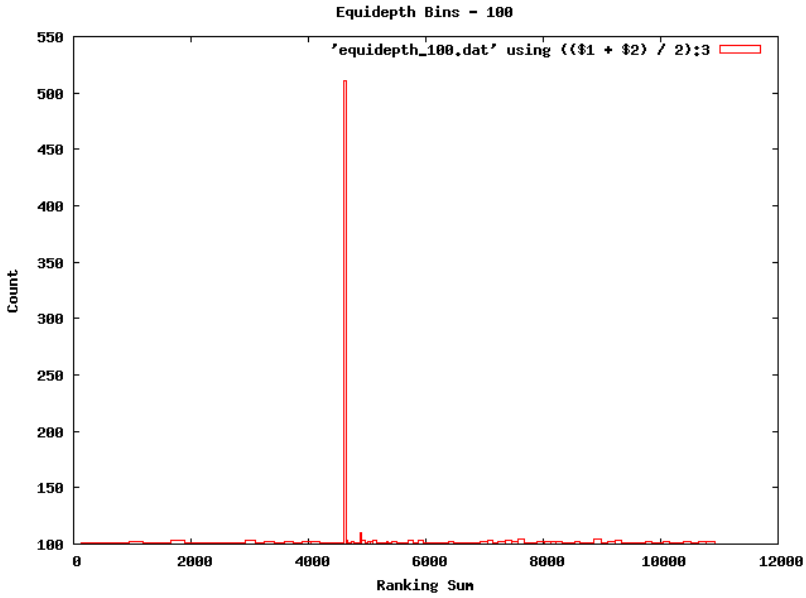

Figure 4.4: Equidepth Binning with Bin Size of 100

similar rankings, or that the range of rankings covered by the bins is too large or too small to provide a good grouping.

For equiwidth binning, we used ranking ranges of 50, 100, 200, 500, and 1000. Figures 4.7, 4.8, 4.9, 4.10, and 4.11 depict these binnings. Just like equidepth binning, this binning strategy is not dynamic to the rankings that it is given. As can be seen by these figures, each bin is equally paritioned, at intervals of 50, 100, 200, 500, or 1000, and so using this binning strategy is essentially the same as saying take all of the individuals with a ranking less than an arbitrarily chosen ranking. 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.

The hierarchical binning strategy avoids these problems by actually considering the distribution of rankings within each bin. For the hierarchical binnings that we created, we used standard deviation cutoffs of 50, 100, 200, and 500. Figures 4.12, 4.13, 4.14, and 4.15 show these binnings. When this binning strategy is used, a tree structure is created, where 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, thus creating two new possible bins. Figure 4.16 contains the tree structure created during the process of performing hierarchical binning with a cutoff of 500.

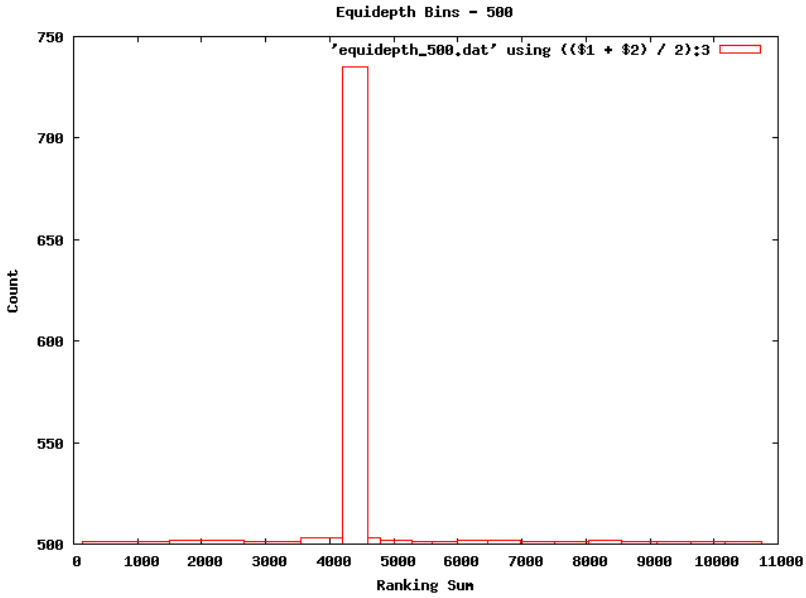

Figure 4.5: Equidepth Binning with Bin Size of 500

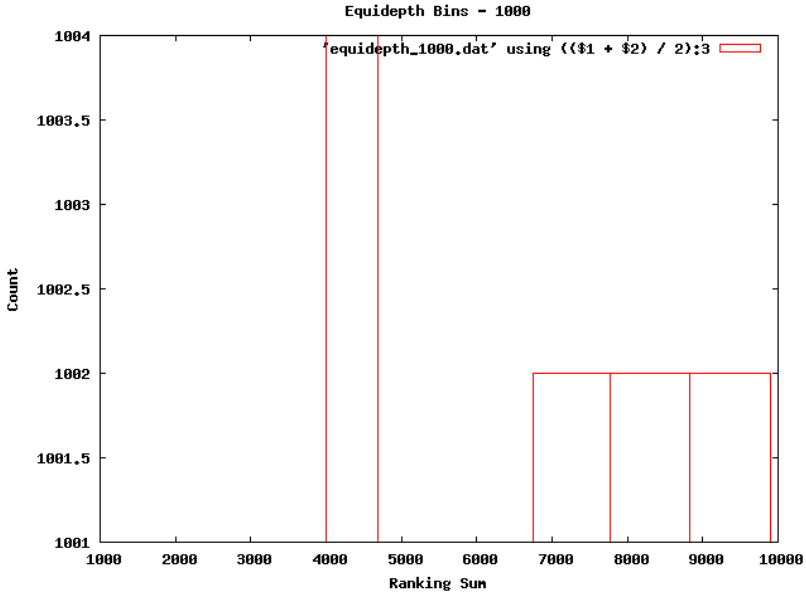

Figure 4.6: Equidepth Binning with Bin Size of 1000

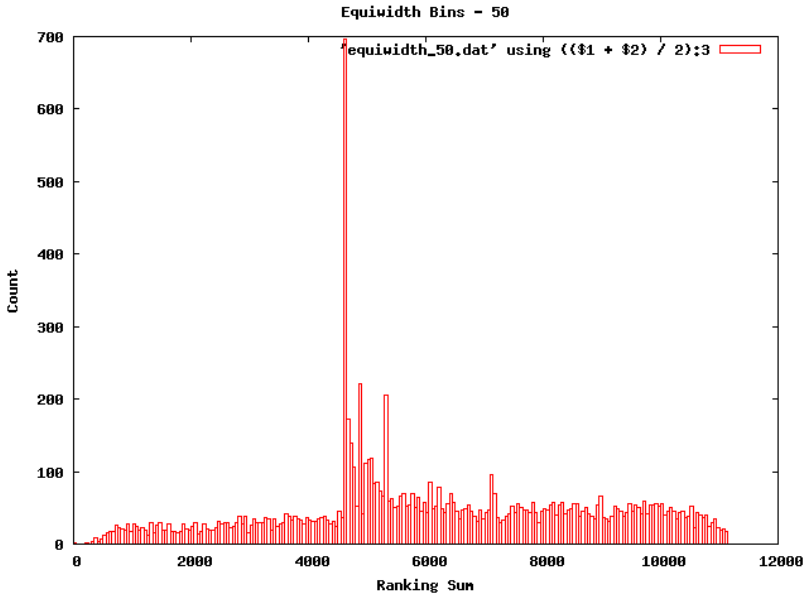

Figure 4.7: Equiwidth Binning with Ranking Range of 50

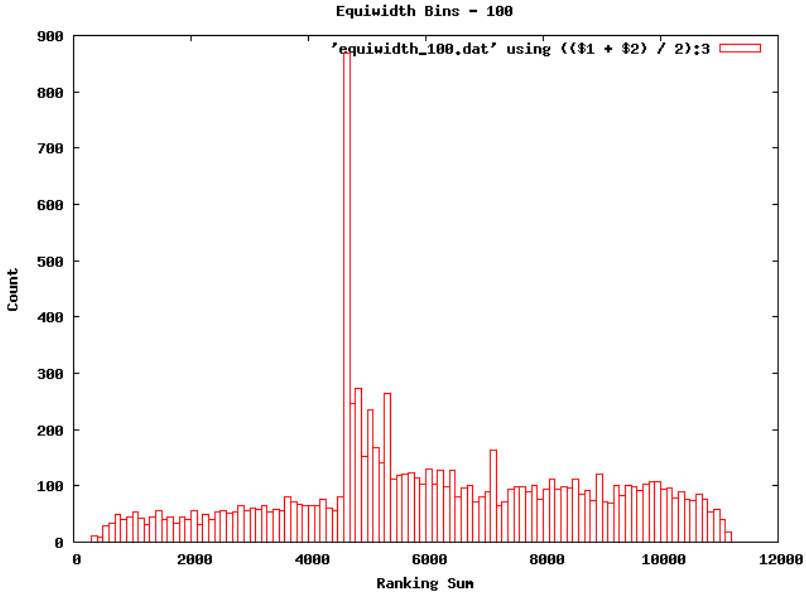

Figure 4.8: Equiwidth Binning with Ranking Range of 100

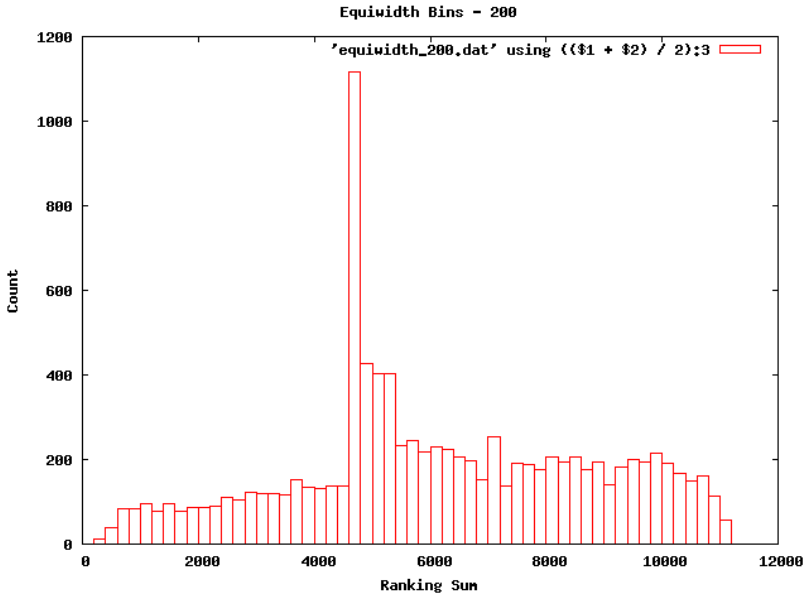

Figure 4.9: Equiwidth Binning with Ranking Range of 200

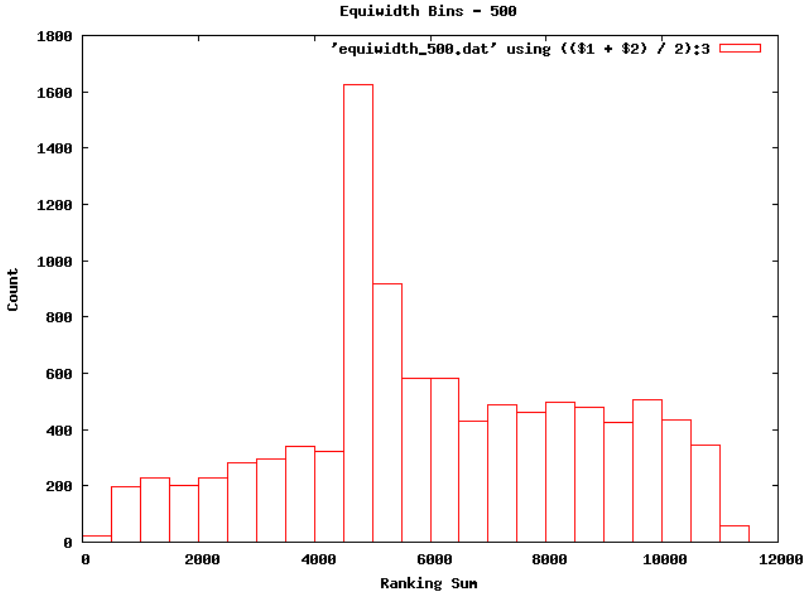

Figure 4.10: Equiwidth Binning with Ranking Range of 500

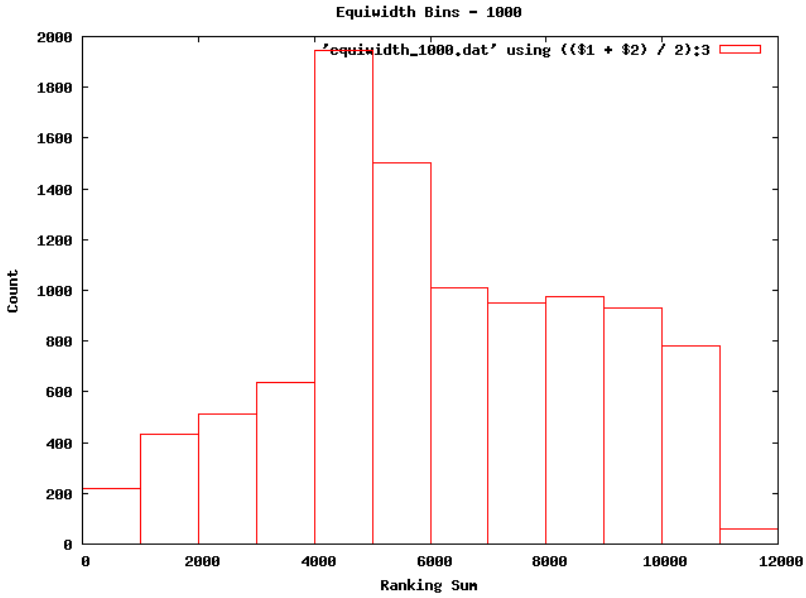

Figure 4.11: Equiwidth Binning with Ranking Range of 1000

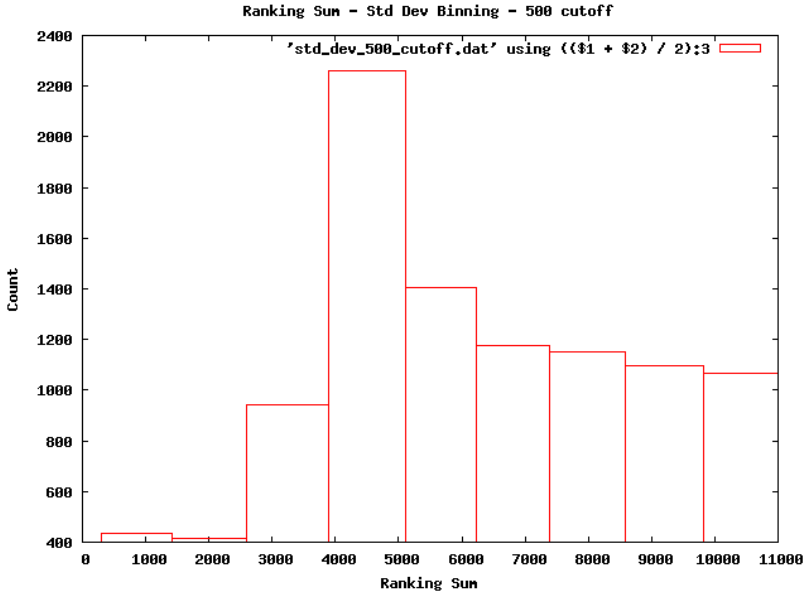

Figure 4.12: Heirarchical Binning with Cutoff of 500

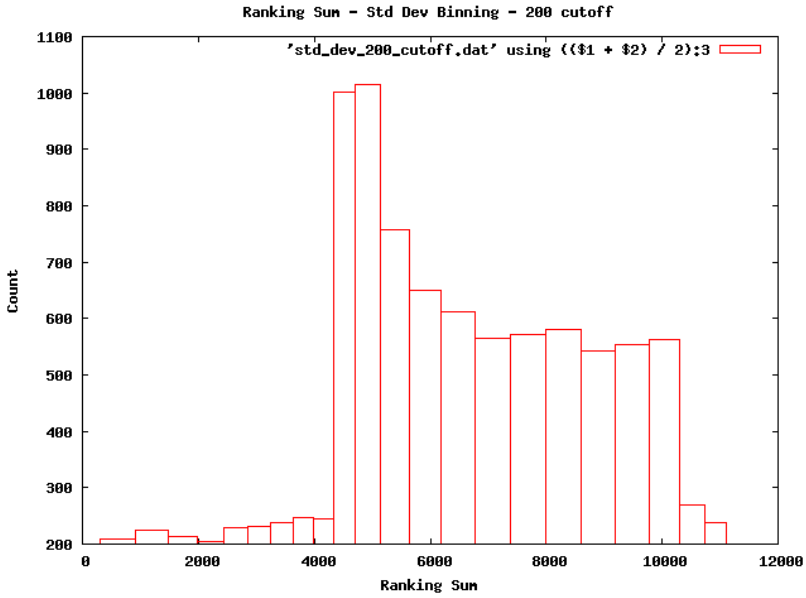

Figure 4.13: Hierarchical Binning with Cutoff of 200

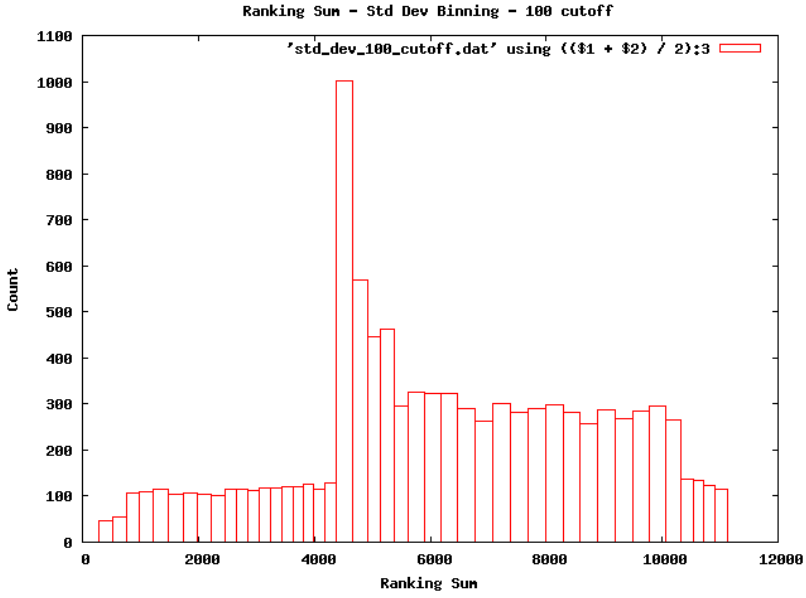

Figure 4.14: Hierarchical Binning with Cutoff of 100

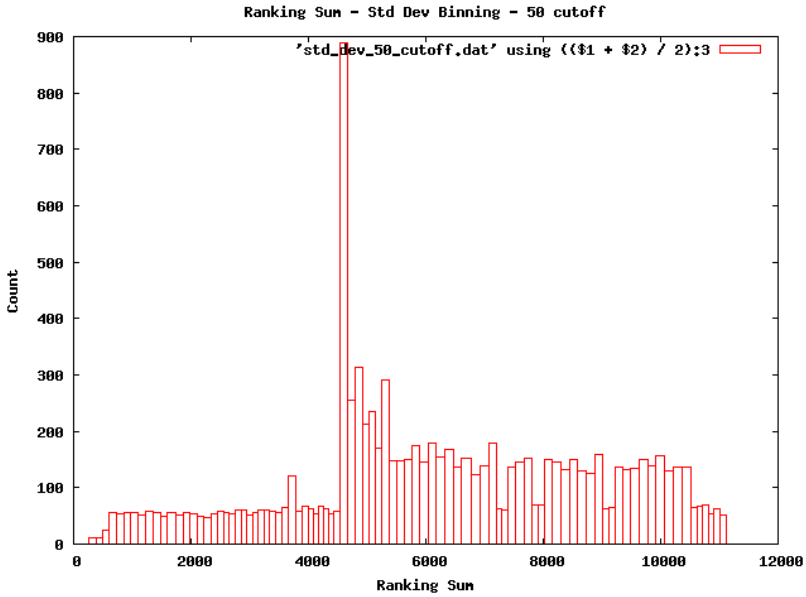

Figure 4.15: Hierarchical Binning with Cutoff of 50

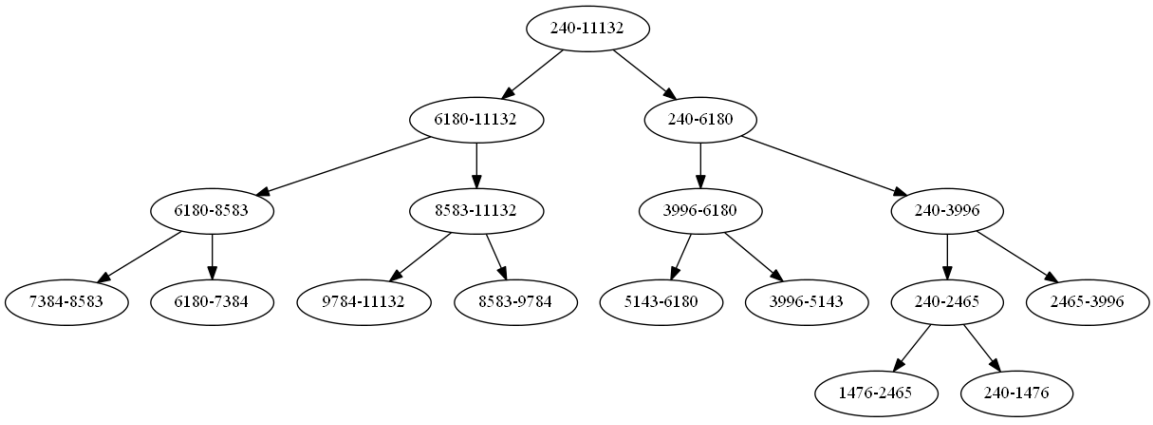

Figure 4.16: Tree Structure of Hierarchical Binning with a Cutoff of 500

Chapter 5

Conclusions

In this thesis, we analyzed those factors that contribute to making an individual in a private data set vulnerable when online public information is considered. We proposed a methodology that can be used to identify the set of most vulnerable individuals with a privately held data set. Then, this methodology was tested using a real world data set.

When the methodology was tested, a small, final set of vulnerable individuals was able to be found within the private data set. The exact composition of this final vulnerable set varied based on the standard deviation cutoff chosen when using hierarchical binning. Using a standard deviation cutoff of 100, there were 210 individuals in the final vulnerable set, with rankings from 240 to 980. When a standard deviation cutoff of 50 was used, there were 46 individuals in the final vulnerable set, with rankings from 240-581.

However, this methodology does more than find the most vulnerable set of individuals within the data set. It also allows companies and organizations to get a sense of the vulnerability of all individuals within the private data set.

5.1 Future Directions

To improve upon this work in the future, the following additions could be made. The methodology could be tested across a greater number of websites. The methodology could be tested on different sets of individuals, and more individuals could be used in these sets. Analysis of vulnerable individuals could be performed in order to discover the attributes that are key in causing vulnerability.

5.2 Acknowledgments

We would like to thank the Census Bureau for access to the public wholesale data set used in this study.

Bibliography

[1] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *Proceedings of the International Conference on World Wide Web*, WWW, pages 351–360, New York, NY, USA, 2010. ACM.

[2] N. Li, T. Li, and S.Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the International Conference on Data Engineering (ICDE)*, April 2007.

[3] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *Proceedings of the International Conference on World Wide Web*, WWW, pages 1145–1146, New York, NY, USA, 2009. ACM.

[4] A. Machanavajjhala, J. Gehrke, and D. Kifer. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[5] A. Ramachandran, L. Singh, E. Porter, and F. Nagle. Exploring re-identification risks in public domains. In *Proceedings of the Conference on Privacy, Security and Trust (to appear)*, 2012.

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

[7] B. Zhou, J. Pei, and W.S. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter*, 10(2):12–22, 2008.