Reidentification of Individuals in Chicago's Homicide Database
A Technical and Legal Study
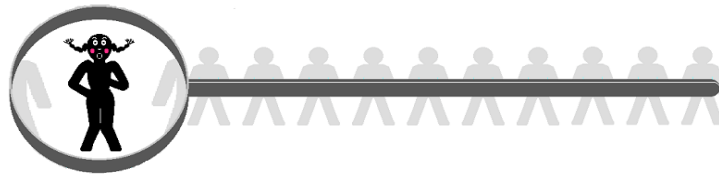
Salvador Ochoa   **Jamie Rasmussen**   **Christine Robson**   **Michael Salib**

Collective address: **reidentify@mit.edu**

Abstract

Many government agencies, hospitals, and other organizations collect personal data of a sensitive nature. Often, these groups would like to release their data for statistical analysis by the scientific community, but do not want to cause the subjects of the data embarrassment or harassment. To resolve this conflict between privacy and progress, data is often deidentified before publication. In short, personally identifying information such as names, home addresses, and social security numbers are stripped from the data. We analyzed one such deidentified data set containing information about Chicago homicide victims over a span of three decades. By comparing the records in the Chicago data set with records in the Social Security Death Index, we were able to associate names with, or reidentify, 35% of the victims. This study details the reidentification method and results, and includes a legal review of U.S. regulations related to reidentification. Based on the findings of our project, we recommend removal of these databases from their online locations, and the establishment of national deidentification regulations.

Table of Contents

Table of Figures

Introduction

We are in an age of rapidly developing technologies that open up possibilities for privacy invasions never before conceived of. With the Internet, the world has been introduced to a new way to compile, exchange, and manipulate data at speeds and volumes heretofore unimagined. Indeed, laws and standards can scarcely keep up with the potentials for privacy invasion.

Our project involves publicly released databases, complied by the United States government for statistical purposes, but disseminated in a manner that allows identification of individuals. In particular, we examined the Chicago Homicide data set, compiled by the Bureau of Justice Statistics and published online by the National Archive of Criminal Justice Data. By combining this data with the Social Security Death Index, also available online, we were able to successfully determine the identity of 35% of the individuals who are supposedly anonymously listed in the database.

In this paper, we will review reidentification theory, paying special note to the work of Professor Latanya Sweeney, of Carnegie Mellon University, and her work with medical databases. We will also describe our methodology for reidentification, including the details of our database matching. A comprehensive analysis of the laws surrounding reidentification is also included. Based on the findings of our project, we will be recommending removal of these databases from their online locations, and the establishment of national deidentification regulations. We conclude the report with both legal and technical recommendations for protection against reidentification.

Reidentification theory

The purpose of this chapter is to provide an introduction to reidentification theory. Later chapters describe a homicide victim reidentification experiment in great detail. This section is intended as a primer for the non-technical reader, explaining many key terms and concepts that will be used throughout this document, so that he/she may fully understand the significance of the project. It is also intended as an overview of the modern trend of increased data collection and sharing, the privacy concerns resulting from such data sharing, and the reasons why reidentification is being done. The technical, informed reader may skip this section without any loss of information.

Key Terms and Concepts

Reidentification concerns manipulating databases to determine the identity of individuals whose information is recorded as records within a deidentified database through data linkage techniques. To best understand this concept, we first define a few terms and then provide a simple example.

A database is a collection of data organized in such a way that a computer program can quickly search for and retrieve desired pieces of information. It is typically stored on magnetic disk or some other secondary storage device, and it is designed to allow for fast and efficient data-processing operations including the storage, retrieval, modification, and deletion of data.

A database can consist of multiple files, each of which is broken down into records. Each record is a complete set of information on a specific entity and is made up of any number of fields, each of which contains information pertaining to one individual aspect or attribute of the entity. For example, a student directory file contains records that may include four fields: a

student name field, an address field, a phone number field, and a major field. Each record may also be considered an *n*-tuple of the *n* different fields that make up the record. A database can be modeled as a simple table where each row corresponds to an individual record and each column corresponds to a field.

Student Name	Address	Phone No.	Major
Alyssa P. Hacker	East Campus	225-5555	Computer Science
Ben Bitdiddle	Next House	225-2222	Electrical Engineering
Joe Law	Baker	225-3333	Political Science
Celine Miles	New House	225-7777	Biology

Figure 1: Table Representation of a Student Directory Database

The above figure depicts a table representation of a student directory database. Each record, or row, contains the directory information for a single student. The record for Ben Bitdiddle is highlighted. Each record is made up of the four fields described earlier, shown as columns. The Address field is highlighted.

The term database is increasingly being used as shorthand for a database management system (DBMS), which is the actual software that is used to perform the data-processing operations mentioned earlier. More formally, a database management system is a collection of programs that enables you to store, modify, and extract information from a database. To be specific, we used PostgreSQL a relational database management system, or RDBMS. These database systems are powerful because they require few assumptions about how data is related or how it will be extracted from the database, and unlike flat database systems, they can work with multiple files.

Requests for information from a database are made in the form of a query, which is a stylized question. For example, the query:

```
SELECT ALL WHERE MAJOR = "POLITICAL SCIENCE"
```

if run on the database in the above figure, would request all records in which the MAJOR field is "Political Science." This query would only result in one value: Joe Law. The set of rules for constructing queries is known as a query language. Although different DBMSs support different query languages, there is a semi-standardized query language called SQL (structured query language), which we used in our project.

Databases, as mentioned, allow for quick retrieval of desired data, or information. This allows for what is now referred to as data mining. Data mining describes finding previously unknown patterns, or relationships in a group of data. In order to support current research in a variety of fields, there has been a tremendous increase in the amount of information that is being collected and stored, so that data mining can produce more results.

Another aspect of databases, which begins to introduce us to the reidentification problem, is the ability to do data linkage. Data linkage refers to combining disparate pieces of entity-specific information to learn more about an entity. That is, a researcher can combine information from different databases about an entity if he/she can match the records. In the figure below,

data linkage of two databases is possible. One database has students' major and GPA information while another has students' biographic information. Each database has student's names, so an administrative official could easily *link* the two databases using students' names to make a single database with all of the students' information.

Name	Age	Sex
Alyssa P. Hacker	21	F
Ben Bitdiddle	21	M
Joe Law	20	M
Celine Miles	19	F

Name	Major	GPA
Alyssa P. Hacker	Computer Science	4.25
Ben Bitdiddle	Electrical Engineering	4.00
Joe Law	Political Science	4.75
Celine Miles	Biology	4.8

Name	Age	Sex	Major	GPA
Alyssa P. Hacker	21	F	Computer Science	4.25
Ben Bitdiddle	21	M	Electrical Engineering	4.00
Joe Law	20	M	Political Science	4.75
Celine Miles	19	F	Biology	4.80

Figure 2: Data Linkage

Although we have been discussing each record as corresponding to an entity, the databases that we are concerned about are those in which each record corresponds to an individual person. In other words, the databases we used in our experiment contain person-specific data, since we are interested in the reidentification of people. Data linkage is important in this respect since it allows for larger profiles.

Growth of Public Data

As a result of the many advancements in computer-related technology in recent years, primary and secondary data storage devices continue to become more affordable. High-speed network connections are also becoming more available to the average consumer as broadband connections such as DSL and cable are increasingly being offered and promoted by Internet service providers.

During recent years, however, as a result of the increased availability of storage devices, society has also been witness to what can only be described as a data explosion. Although we recognize that we live in the Information Age, what many do not realize is that much of the information that is being collected today is about individuals. Latanya Sweeney, one of the trailblazers in the field of reidentification research and theory, has described in her thesis on reidentification that "there has been tremendous growth in the collection of information being collected on individuals and this growth is related to access to inexpensive computers with large

storage capacities."[1] She also asserts that because the affordability of these systems will only increase in the years ahead, "the trend in collecting increasing amounts of information is expected to continue. As a result, many details in the lives of people are being documented in databases somewhere."

Her research has led her to find three major trends with regard to data collection: (1) "collect more;" (2) "collect specifically;" and, (3)"collect if you can."[2] As an example of the collect more trend, she describes how birth records moved from having only seven to fifteen fields per live birth at the beginning of the twentieth century, to about 25 fields in later years, but jumping to over 100 fields per live births as the availability and use of electronic equipment in hospitals and clinics has increased in the latter part of the century.[3] By "collect specifically," she means that instead of collecting tabular information, many entities are now collecting person-specific information. She lists supermarkets as an example; they, using the now familiar loyalty, or saver cards, can collect information about clients' purchases. She also points to the fact that many entities are now collecting information simply because it has now become possible for them to do so. These include immunization record databases for example.

Sweeney, using what she refers to as the global disk storage per person factor, or DSP, attempts to characterize the growth in person-specific data. By dividing the amount of disk storage space sold worldwide in a given year and dividing by the world population at that time, she obtains the GDSP, which she claims is "a crude measure of how much disk storage could possibly be used to collect person-specific data on the world population." The figure below depicts her estimates and illustrates how the GDSP value is growing dramatically.

	1983	1996	2000
GDSP (MB/person)	0.02	28	472

Figure 3: GDSP Over Time

Privacy Concerns

The amount of personal information collected should be enough to raise privacy concerns. However, the real problems arise when we begin to consider the availability of all of this information. As mentioned before, network connectivity is becoming ubiquitous; high-bandwidth connections especially are becoming popular as they become more affordable. Over the years, there has been a noticeable trend in making more databases available online, as well as offline, because of the ease of data transfer that it allows. Some states, such as Texas, have their birth and death registries online, medical data, including hospital discharge data, is readily available, and even health and criminal records are accessible.

The dramatic increase in databases available online is attributable to researchers' interest in sharing data so that anyone can use the data to aid in their own studies. Some databases may be made available for more superficial reasons such as profit in the case of marketing databases. Along with what appear to be "innocent" databases, there is a great quantity of databases that contain personal, private information. These databases may include health records, police reports, etc. For example, health records can contain abortion records, which many women who have had abortions would surely not want to be made public.

---

[1] Sweeney, pg. 20
[2] Sweeney, pg. 41
[3] Sweeney, pgs. 6-11

Access Policies

The data holders, often the data collectors themselves, recognize that much of the information they are protecting may be personal, but they are also influenced by the fact that the data they hold may be the key for some important discovery. They are then forced to choose an access policy for their data. Latanya Sweeney also addresses this point in her PhD thesis. She states that there are four basic access policies: (1) private, meaning "insiders only;" (2) semi-private, or "limited access;" (3) semi-public, or "deniable access;" and, (4) public, meaning "no restrictions." [4]

A private database, essentially, is one that is not shared with anyone. Usually, only the data collectors themselves have access to the data. Databases that are semi-private are fairly similar in that they are shared with only a very select few. There is usually some type of rigorous review process before access is granted. For databases that are ruled by either of these access policies the privacy concern is small. The private information is not being shared and data holders probably obtained their subjects' information directly from them.

The privacy concern is more explicit in databases that are controlled by public or semi-public access policies. Semi-public databases are available to a great number of people. The number of people or entities denied access is very small compared to how many are granted access. Public databases have absolutely no restrictions and are available to anyone who requests access. For databases that contain personal information, but adhere to either of these access policies, the protection of the privacy of their subjects should be paramount. Subjects' privacy can only be assured by anonymizing the released data.

Usefulness of Data

However, data holders are faced with an additional dilemma – as data is made more anonymous, it becomes less useful. That is, there is an inverse relationship between the anonymity and usefulness of data. For example, a researcher can make much more use of a fully identified database, one that leaves all personally identifiable information, such as name and address, than with purely aggregate statistics. R.J.A Little states that methods to anonymize data "are known to reduce the analytic validity of files,"[5] because, as Sweeney explains, "any attempt to provide some anonymity protection, no matter how minimal, involves modifying the data and thereby distorting its contents."[6] Thus, from a researcher's point of view, no modification of the data is desirable.

The data holder must then determine to what extent the data must be anonymized. This, if possible, can be done on a per-release basis, evaluating the subjects' privacy against a recipient's purported need for the information. Sweeney suggests that there are cases where the privacy of the data greatly outweighs any possible need by outsiders. This is the case for classified government data, or a company's employment records (do not want to give away the names of their high performers). In this case, all information is completely suppressed, i.e. no data is released. At the other extreme, there is the case where the recipient's need overshadows any privacy concerns. In this case, the data is released with no modifications and all subjects completely identified. An example of this case is a public health official's request for health records.

---

[4] Sweeney, pg 42
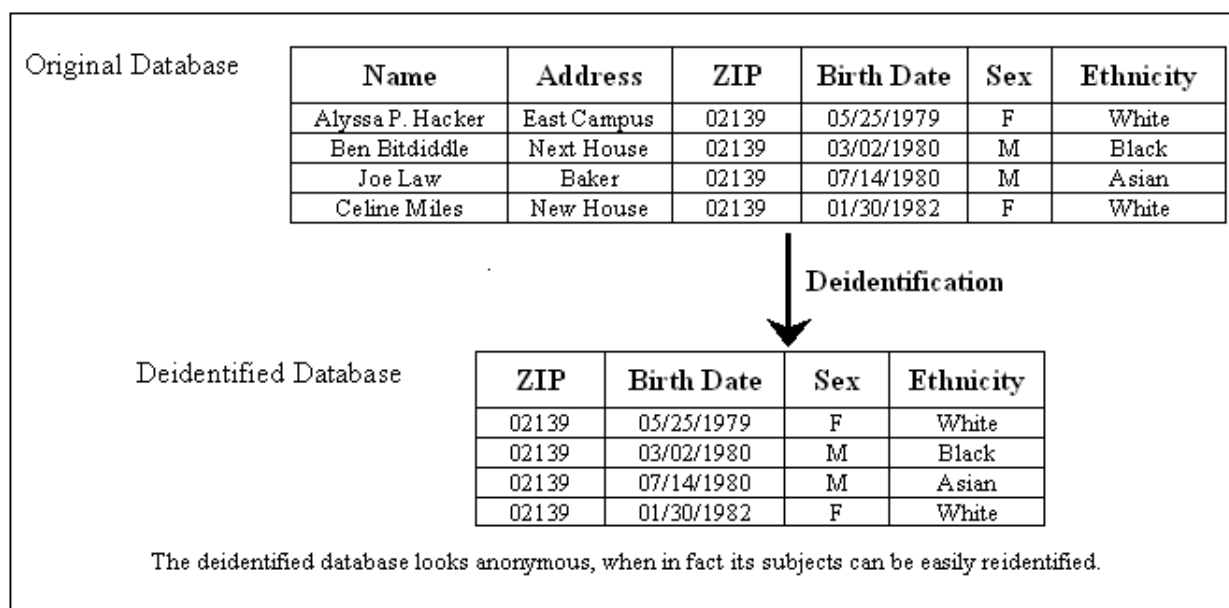[5] Little, R. J. A. (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics*, 9, 407-426.
[6] Sweeney, pg 31

In between these two cases, however, there is an extremely wide band. Sweeney describes it as a continuum, with the two cases mentioned as the endpoints. She argues that most cases fall somewhere in this continuum and that the problem then becomes that data holders release data that is too distorted in an effort to anonymize, or is easily reidentifiable. That is, they do not achieve the "optimal release of data" – a release of data that is practically useful yet is minimally invasive to subjects' privacy.[7]

Deidentification

Since the focus of this document is on subjects' privacy, we direct our attention to the case where a release of personal data is not completely anonymous. Investigators (i.e. Sweeney, other reidentification researchers, and we) have found that many database releases are made public under the mistaken assumption that simply removing explicit identifiers from the databases' records makes them anonymous. Explicit identifiers are data fields that contain personally identifiable information; Sweeney defines explicit identifiers as, "a set of data elements, such as {name, address}, for which there exists a direct communication method where with no additional information, the designated person could be directly and uniquely contacted."[8] Although they do not fit the definition of explicit identifiers, Social Security numbers are also usually removed from these supposedly anonymous databases because they are in such widespread use and their holders can be identified easily.

The removal of all explicit identifiers from a database is termed deidentification. It is important to note, however, that although a deidentified database may appear anonymous (see Figure below), it certainly is not. Deidentification is a misnomer, since deidentified data is not equivalent to anonymous data. We define deidentified data simply as data that has undergone deidentification – explicit identifiers have been removed, generalized, or replaced with fictitious data – whereas, anonymous data is data that cannot be manipulated to reidentify the subject of the data.

Original Database

Name	Address	ZIP	Birth Date	Sex	Ethnicity
Alyssa P. Hacker	East Campus	02139	05/25/1979	F	White
Ben Bitdiddle	Next House	02139	03/02/1980	M	Black
Joe Law	Baker	02139	07/14/1980	M	Asian
Celine Miles	New House	02139	01/30/1982	F	White

Deidentification

Deidentified Database

ZIP	Birth Date	Sex	Ethnicity
02139	05/25/1979	F	White
02139	03/02/1980	M	Black
02139	07/14/1980	M	Asian
02139	01/30/1982	F	White

The deidentified database looks anonymous, when in fact its subjects can be easily reidentified.

---

[7] Sweeney, pg. 31
[8] Sweeney, pg. 14

Reidentification

The distinction between deidentified data and anonymous data thus lies in the ability to subject the data to reidentification. Reidentification is the discovery, or determination, of the identity of the individuals who are the subjects of a study through data linkage techniques. It only applies to reidentification of subjects when the data holders have attempted to deidentify them in some manner. That is, a fully identified database cannot be said to undergo reidentification.

Within the vast amount of personal information that is being collected as part of the 'data explosion,' there is personal data that is extremely private for the subjects, data that they would not be connected to publicly. A later section provides a few examples of these data sets. In most cases, the data is only publicly available because the subjects have been assured of their privacy – they have been assured that the data will be anonymous. Reidentification, then, raises grave privacy concerns because of the simple fact that it voids the attempts of many researchers to protect the privacy they have guaranteed to their subjects. It is a tool for invasion of privacy, and it will be increasingly possible for reidentification to take place, with much greater ease and by a greater number of people, as the amount of data available continues to grow.

Reidentification is a relatively simple concept. It makes use of what Latanya Sweeney terms 'quasi-identifiers.' A quasi-identifier is "a set of data elements in entity-specific data that in combination associates uniquely or almost uniquely to an entity and therefore can serve as a means of directly or indirectly recognizing the specific entity that is the subject of the data."[9] It is a combination of characteristics that, combined, can act as a unique or near-unique identifier in the absence of explicit identifiers. For example the set consisting of a person's home ZIP code, gender, and birth date does not contain any explicit identifiers, but can be a quasi-identifier since this set can uniquely identify a large percentage of the population. Sweeney found that this quasi-identifier made 87% of the population in the United States unique and identifiable; birth date and full ZIP code alone makes 97% of the Cambridge, Massachusetts population identifiable.[10] Basically, a few characteristics can make a person unique.

Using an exhaustive control data set, one can determine a quasi-identifier that can uniquely identify the largest number of individuals. An exhaustive control data set is a data set that contains personal information, including explicit identifiers, about a large percentage of the population from which the subjects of a deidentified database are drawn. For example, voter registration lists contain information such as name, address, ZIP code, birth date, and gender of each voter, in addition to party affiliation and date registered, about a large percentage of adults for specific areas. Thus, they often make excellent control data sets. It is using the Cambridge voter list that Sweeney found that 97% of its population was uniquely identifiable using certain data. It is through the analysis of the voter list as the control data set that she was able to find that the quasi-identifier that would give this high percentage was {full ZIP, birth date}. As the amount of information given in the control data set increases – has more, specific fields – the better a quasi-identifier will be. It is also important to note that a control data set does not have to be public. Companies can use their own employee records as a control database – it contains information about all of its employees!

---

[9] Sweeney, pg. 17
[10] Sweeney, pgs. 49-50

A data investigator – anyone with data storage space, (network) access, database software (a DBMS), and interest – can then use a good quasi-identifier to match a large number of the subjects of a deidentified database to the individuals named in the control database. That is, he/she will use data linkage techniques to match the private information in the deidentified database to an identity in the control database using the shared quasi-identifier information as the linking data. Figure 5 illustrates this process.

Deidentified Database			Control Database		
Subject	Personal Information	**Shared** Quasi-identifier {ZIP, birth date, sex}	Name	Street Address	Party Affiliation
1	Illness 12	{02139, 08/28/52, M}	Bill Robinson	1 Main Street	Democrat
2	Illness 1	{02139, 02/14/62, F}	Marsha Wilkins	15 Broadway	Republican
…	…	…	…	…	…

Figure 5: Linking a Deidentified Database with a Control Database

An Example

This subsection provides a simple, complete example of the reidentification process. We include it in order to better explain the procedure and illustrate how easily anyone can perform reidentification of subjects.

The example-deidentified database contains information about subjects who have sexually transmitted diseases (STD). The subjects considered their diagnosis private information and did not want to be identified as having been diagnosed with an STD. The data collectors guaranteed them that their identities would not be made public when they released their patient data. They thus deidentified their data, believing it was rendered anonymous, before releasing it. Figure 6 depicts the data that they made public.

ZIP	Birth Date	Sex	Diagnosis	Diagnosis Date
02139	05/25/1979	F	AIDS	01/21/01
02139	03/02/1980	M	Flu	11/15/00
02139	07/14/1980	M	HIV	03/28/01
02139	01/30/1982	F	Neuroblastoma	07/08/99

Figure 6: Deidentified Private Information Made Public

Since all of the subjects live in the same area, as specified by the ZIP code field, and are of voting age, a suitable control database would be the voter registration list for their area. It is depicted in Figure 7 below.

Name	Street Address	ZIP	Birth Date	Sex	Date Registered	Party Affiliation
Alyssa P. Hacker	15 Main St.	02139	05/25/1979	F	06/15/97	Democrat
Ben Bitdiddle	68 Broadway	02139	03/02/1980	M	03/03/98	Republican
Joe Law	86 Central Ave.	02139	07/14/1980	M	10/28/80	Democrat
Celine Miles	21 South St.	02139	01/30/1982	F	02/02/00	Republican
…	…	…	…	…	…	…

Figure 7: A Control Database - Voter Registration List

A data investigator, looking at the two data sets, sees that both contain ZIP, birth date and sex information. This set of data can then be used as a quasi-identifier. Figure 8 illustrates this overlap in data.
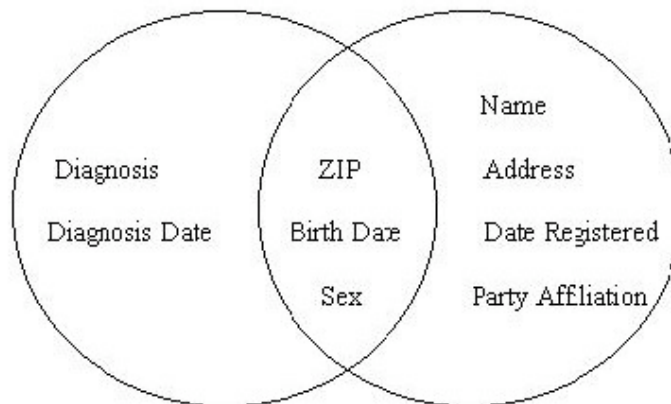

Figure 8: Overlap in Data in the Two Data Sets

The data investigator can then attempt to match the subjects in the deidentified patient database with the individuals in the control database using the quasi-identifier as the basis for linkage of diagnosis to identity. The results of this linkage are shown in Figure 9.

Name	Street Address	ZIP	Birth Date	Sex	Diagnosis	Diagnosis Date
Alyssa P. Hacker	15 Main St.	02139	05/25/1979	F	AIDS	01/21/01
Ben Bitdiddle	68 Broadway	02139	03/02/1980	M	Flu	11/15/00
Joe Law	86 Central Ave.	02139	07/14/1980	M	HIV	03/28/01
Celine Miles	21 South St.	02139	01/30/1982	F	Neuroblastoma	07/08/99

Figure 9: A Reidentified Data Set

Although all of the subjects in our deidentified database were reidentified, this is not always the case. Sometimes the control data set does not contain a match, or contains more than one. It might still be possible to positively reidentify the subjects who fall into these categories, however, by looking more closely at other data fields.

Reasons for Reidentification

The reidentification example illustrated how easy it is to do reidentification. However, we are left with the question: Who would reidentify? In fact, there are many people or entities that would be interested in reidentification of private, deidentified data subjects. This section provides a few reasons for which they may use reidentification.

Scientific Research

Scientific research is one of the main reasons much of the data available is ever collected and shared. As scientists form and test their hypotheses using deidentified data sets, they may find that they need additional information about the subjects in order to complete their research.

They may need information that is simply more useful than the deidentified information they have. They wish to reidentify the subjects so that they can build a larger profile on each of the subjects, or for a select few.

For example, a medical researcher studying health issues may have a deidentified data set containing certain, general characteristics about some individuals' medical histories. He finds that a few subjects have data that is unusual, or interesting in some way. If he could identify and contact those subjects in order to obtain more information about them, then it would be greatly beneficial for his research. Although this seems innocent enough, one must consider that some individuals may not want to be contacted or even have their information linked to them by anyone other than their doctor.

Investigative Reporting

Reidentification can be used for many different types of investigative reporting. Reporters may try to link personal information contained in deidentified data sets to celebrities or public officials and report the information gathered about them to the public at large. Sweeney, in her thesis, provides an event that can be used as an example. She writes, "In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. GIC collected de-identified patient-specific data with nearly one hundred fields of information per encounter along the lines of the fields discussed in the NAHDO list for approximately 135,000 state employees and their families. Because the data were believed to be anonymous, GIC gave a copy of the data to researchers and sold a copy to industry."[11] Among the data subjects were well-known, high-ranking officials, including the governor. Obviously, if his personal medical data could be reidentified, then the press could quickly make his private medical information public. Actually, Sweeney writes that the governor's data could be uniquely identified using only his birth date, sex, and five-digit ZIP code.[12]

Marketing

Marketing provides the impetus for much of the increased data collection characteristic of recent years. Marketers want to build the largest profiles about consumers as possible in order to be able to do greater direct marketing. This would allow them to increase profits by narrowing the amount of people the market certain products to, while, at the same time, increasing the probability of success for each direct marketing target.

Just recently, Doubleclick, Inc., an online marketing firm that tracks users browsing habits, sought to reidentify many of its subjects by buying a consumer database. Although it was thwarted by its own privacy policy, the privacy danger was real. Doubleclick would have been in the position to identify individuals with their browsing habits and be able to sell this information to other product or service providers.

Blackmail

Blackmail is an interesting motive for doing reidentification. Although it does not seem apparent that reidentification would be useful for reidentifying information for a particular, specific individual, there is the possibility of reidentifying celebrities, public officials, or anyone

---

[11] Sweeney, Pg. 50
[12] Sweeney, Pg. 50

else with very personal information that a malicious data investigator may threaten to make public unless the reidentified individual meets some demand.

There are already public databases that contain arrest data for certain police districts. If all such information were made available, then a data investigator could surely reidentify well-known individuals with their arrest record. They could then attempt to blackmail the individuals by threatening to make their record public.

Insurance

Health and life insurance companies have a very real motive for attempting to do reidentification. This may be another reason that the medical field has been attempting to bring attention to the reidentification issue. These insurance companies can attempt to reidentify individuals in deidentified hospital discharge data, which is widely available, or other patient data in order to collect a greater amount of information regarding individuals' medical histories. They can then use this greater amount of information to deny certain individuals any type of insurance policy.

Political Action

Yet another reason for attempting to do reidentification is for political motives. Recently, there was a case where an anti-abortion group posted the names and addresses of doctors that conducted abortion procedures on women. As doctors were killed, their names would be crossed out on this list. Now, however, with reidentification, it would be possible to identify the actual women who have had abortions. This is a frightening possibility since public disclosure of their identities might subject them to harassment, danger, as well as discourage other women from seeking abortion.

Reidentification of women who have had abortions would be possible because hospitals and clinics collect and share a great amount of patient data. Within this data is also information regarding procedures performed, including abortions. A political activist could then separate out the subjects who are indicated as having had abortions and try to reidentify them.

Database Selection Criteria

Upon deciding we wanted to conduct reidentification experiments, our first step was to locate a database that contained deidentified data. In addition, the candidate data set had to have certain properties in order to be most useful to us. Specifically, it had to be small, since we had never done this before and wanted to start by working on a tractable problem that could be analyzed quickly without expending a great deal of time or computational resources. We also wanted the candidate database to contain incriminating or embarrassing information about the individuals that had been deidentified. After all, there is little point in expending a great deal of energy to reidentify people only to discover trivia. Trivial information about individuals is much less likely to be well-protected using strong deidentification techniques, and as a result, is unlikely to be representative of the challenges involved in reidentifying important data (like health care information).

Another criteria for our candidate data set was that it had to be easy to verify. From the beginning, we felt it was important to not only make successful reidentifications, but to have some method of verifying the legitimacy of those matches. While such considerations are significantly less important in a commercial setting because the cost of being wrong is so low, we did not conduct our experiments in such an environment. In addition, we wanted to focus on

an area of that had not been as widely explored as medical data. In particular, Latanya Sweeney has written a great deal on that subject and we feel there is little we could contribute in that area. Finally, we required that the candidate data set be available to the public at large for free or for a nominal fee. While many large corporations and government entities maintain large deidentified data sets for internal use, we felt the best way to illustrate the threat from reidentification would be to only work from publicly sources.

We eventually settled on the Chicago Homicide Data set since it met all the criteria listed above. It was small, contained a wealth of embarrassing information, and was freely available to the general public. Additionally, it was in an area that had not received the strict privacy analysis and regulatory burdens that health care data had recently undergone. The data set contained enough personally identifying fields to make reidentification at least plausible and initially appeared to be easily verifiable, although this later turned out to not be the case.

Chicago Homicide Data

The Chicago Homicide Database consists of an exhaustive record of all murders that occurred in Chicago, Illinois from 1965 to 1995. This data was recovered from police logs and includes detailed information on both offenders and victims. The data set includes information on approximately 23,000 victims and 26,000 offenders.

Structure

The Chicago Homicide data set was most useful to us in that it included fields with which to reidentify the victims listed. In particular, the fields describing the day, month, and year of death, as well as the victim's age, gender, and race were invaluable. Also beneficial were fields describing the location of the homicide, both with respect to the victim (home, work, etc.) and in terms of census tract numbers. The data set included a wealth of other fields that might prove embarrassing or incriminating for victims and their families. This included fields such as the relationship between the victim and the offender, the reason for the homicide, previous criminal histories of the victim and offender, as well cause and motivation for the homicide. In addition, the data set includes flags indicating whether the murder involved drugs, child abuse, gang violence, or domestic abuse.

Source:	Illinois Criminal Justice Information Authority
Size:	4.8 MB
Dates Covered:	1965-1995 (only 1982-1995 have death date, location code)
Record Count:	23,817 victims (data on offenders available in separate database)
Covers:	Chicago
Cost:	Free

Figure 10: Chicago Homicide Victims Data Set at a Glance

deathyr	deathmon	deathdte	numvic	vicsex	vicage	vicrace	centract	childabs	gang
95	12	15	1	1	19	2	3802	0	0
95	12	16	1	1	21	2	4211	0	1
95	12	15	1	1	49	2	2702	0	0
95	12	16	1	1	38	1	2304	0	1
95	12	18	1	1	24	1	4004	0	0
95	12	18	1	1	50	2	4604	0	0
95	12	20	1	2	69	2	2817	0	0
95	12	20	1	1	22	3	2427	0	0
95	12	20	1	1	18	2	4207	0	1
95	12	17	1	1	1	1	6107	1	0
95	12	21	1	1	19	3	3114	0	0
95	12	23	1	1	15	2	2507	0	1
95	12	23	1	1	39	2	1914	0	1
95	12	24	1	1	44	3	309	0	0
95	12	24	2	1	53	2	6708	0	0
95	12	24	2	1	19	2	6708	0	0
95	12	26	1	1	26	3	5203	0	0
95	12	26	1	1	21	3	5203	0	1
95	12	27	1	1	26	2	2606	0	0
95	12	27	1	2	46	2	106	0	0
95	12	27	1	2	36	2	4313	0	0
95	12	28	1	1	25	2	3817	0	0
95	12	28	1	2	16	2	2909	0	0

Figure 11: Sample of the Chicago Homicide Victims Data Set

Because the information in the data set was collected over the period of 30 years, it does not provide a complete picture of Chicago homicides. Some fields were added to the data set well after it was started; for example, victims' ages are not reported before 1982. This reduces the number of people that could possibly be reidentified to about 10,000. In addition, some fields reference time varying information. For example, each victim record indicates the police district in which the murder occurred. Unfortunately, the boundaries between police districts in Chicago have changed considerably in the last 30 years as new districts were created and existing districts' boundaries were reorganized. This complicates geographical analysis of the data using police districts considerably. Finally, because young males are disproportionately likely to be involved in homicides, this data set is skewed in the sense that young males are over-represented.

Statistics

Before attempting to reidentify victims in the Chicago data set, we performed a preliminary analysis to determine the likelihood of finding unique matches using the Chicago Homicide data set. We focused on measuring the number of unique instances of {death year, death month, death day, victim age} tuples in the data set. Our analysis found that 93.5% of the records are uniquely identified by this tuple in the homicide data set, while 6.2% of the records match one other record based upon this tuple, 0.22% match two other records based upon this tuple, and 0.073% match three other records. This analysis only covers uniqueness in the Homicide data set itself, not in an exhaustive register, so it can only give an upper bound, or best-case scenario for our reidentification.

Due to the age skew concern mentioned above and revisited in the SSDI chapter, we decided to group the homicides in the Chicago data set by age. Figure 12 shows the age distribution of homicides nationwide, while Figure 13 shows the age distribution of homicides in Chicago. From the similarity of the graphs we can conclude that the Chicago data is not atypical with regard to age distribution.
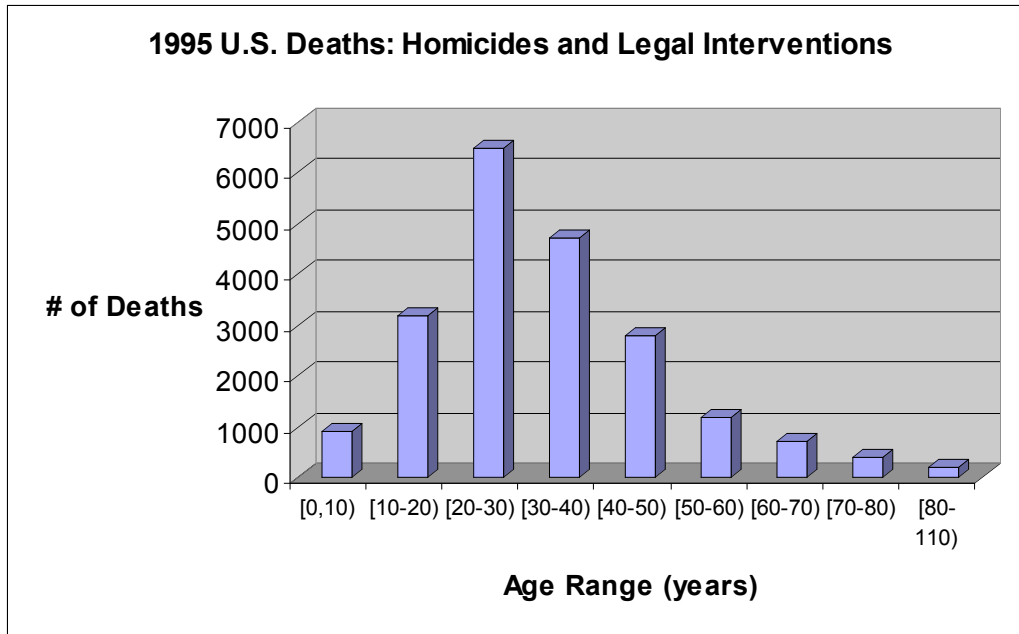
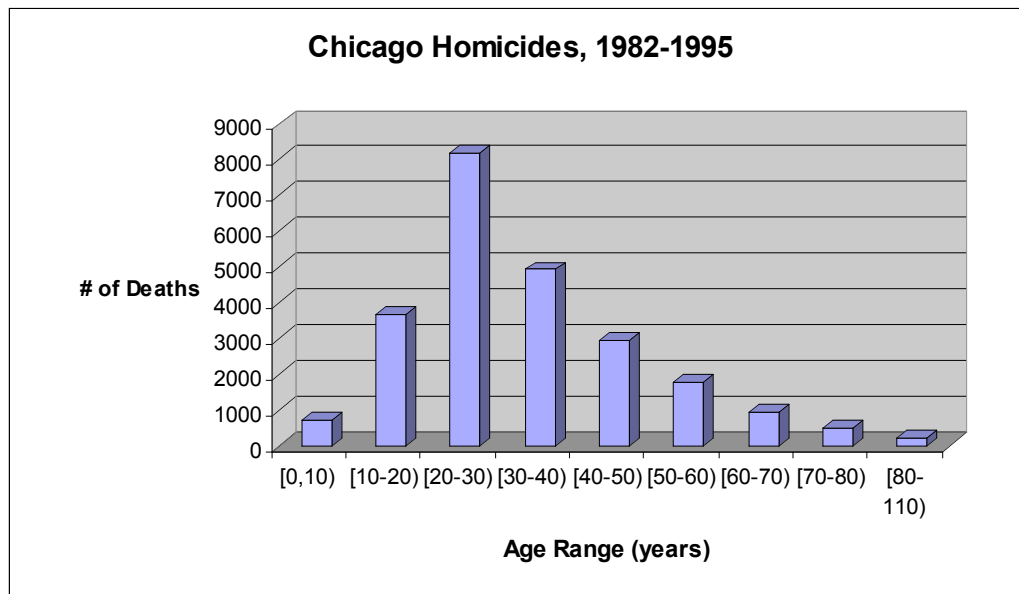**Figure 12: U.S. Homicides and Legal Interventions by Age Range, 1995**

Figure 13: Chicago Homicides by Age Range, 1982-1995

SSDI

The Social Security Death Index (SSDI) is the common name of electronic interfaces to copies of the Social Security Administration's Death Master File (DMF). The DMF contains about 65 million records, one for each death that was reported to the SSA. Although it contains records of people born as early as 1800, close to 98% of the entire data set is individuals who died after 1962, which is the year the SSA began keeping computerized records.

The Social Security Administration sells the DMF to the public in a tape format or on CD-ROM through the U.S. Department of Commerce, National Technical Information Service (NTIS). The cost is $1,725 for a one-time order of the entire data set and $6,900 for the entire file with monthly updates. As the SSA has never provided Internet access to the DMF, some of the purchasers have created free searchable Internet indices, renaming the database the Social Security Death Index. Two such purchasers are RootsWeb.com and Ancestry.com. We decided to use RootsWeb for our research, as it has an easily exploitable interface and spry servers.

We decided that the SSDI would be our control data set, so we downloaded in bulk all of the records for which the last known residence was Chicago, IL and death occurred between the years 1982 and 1995. (See Appendix A for a technical overview of how this was accomplished.)

Structure

Each record in the SSDI corresponds to a deceased person. There are fields for the individual's last name, first name, date of birth, date of death, zip code of last residence, zip code of last payment, SSN, and the state that issued that person's SSN. For formatting reasons, Figure 14 has been edited to remove the zip code of last payment. Some important things to note about the sample records in the figure:

1. Gender is not explicitly specified, but can usually be guessed from the first name, e.g. Mary, Eva, and Violet are probably female, while Edward, Allen, and Andrew are probably male.
2. Ethnicity is not explicitly specified, but could possibly be guessed from the last name. e.g. Perez and Garcia are common Hispanic names. This connection is less assured than the gender connection; for many reasons, an individual's last name may not correspond to his actual ethnicity. (Because this connection is so tenuous, and reliable statistics are not readily available, we never considered race or ethnic codes when reidentifying.)
3. Sometimes the fields in the SSDI are missing or incomplete. Specifically, note that two of the records shown in Figure 14 list only the month of death and not the day. Furthermore, one record has only the first initial and not the entire first name.

```
NAME              BIRTH         DEATH          LAST ADDR.        SSN       ISSUED
MARY PAQUEREAU    01 Jun 1905   21 Apr 1995    60620     002-09-5332      New Hampshire
EVA PLATZMAN      28 Jun 1923   01 Jul 1995    60643     002-14-9381      New Hampshire
LONNIE WILLIAMS   24 May 1926   28 Feb 1995    60619     133-22-5678      New York
YUSUF MOHAMMAD    25 May 1929   Sep 1995       60625     110-42-3155      New York
VIOLET HASKINS    12 May 1906   13 Sep 1995    60618     318-05-0617      Illinois
EDWARD RENK       03 Jun 1905   07 Aug 1995    60638     318-05-1678      Illinois
S KHOMUT          03 Dec 1934   Dec 1995       60640     361-86-0727      Illinois
ALLEN POOR JR     15 May 1958   21 Nov 1995    60660     577-88-6200      District of Columbia
RAUL PEREZ        10 Mar 1942   19 Sep 1995    60620     580-03-7340      Virgin Islands or Puerto Rico
JUSTINO HORTA     11 Aug 1926   19 Feb 1995    60647     581-09-3276      Puerto Rico
ANDREW TAFT       07 Jan 1910   15 Feb 1995    60657     700-18-5134      Long-time or retired railroad workers
RAFAEL GARCIA     20 Jun 1918   03 May 1995    60634     728-09-0476      Long-time or retired railroad workers
```

Figure 14: Sample of Social Security Death Index

Statistics

We were initially worried about the suitability of the SSDI for our reidentification efforts. Specifically, we wanted to know how complete the records were, and if there was any appreciable age skew due to the method of collection. (Our conjecture was that deaths might only be reported for those who would have received death benefits.) We discovered that the SSDI is fairly complete. As shown in Figure 15, the SSDI contains about 92.5% of the total United States deaths recorded by the U.S. Census Bureau for the years 1994-1996.
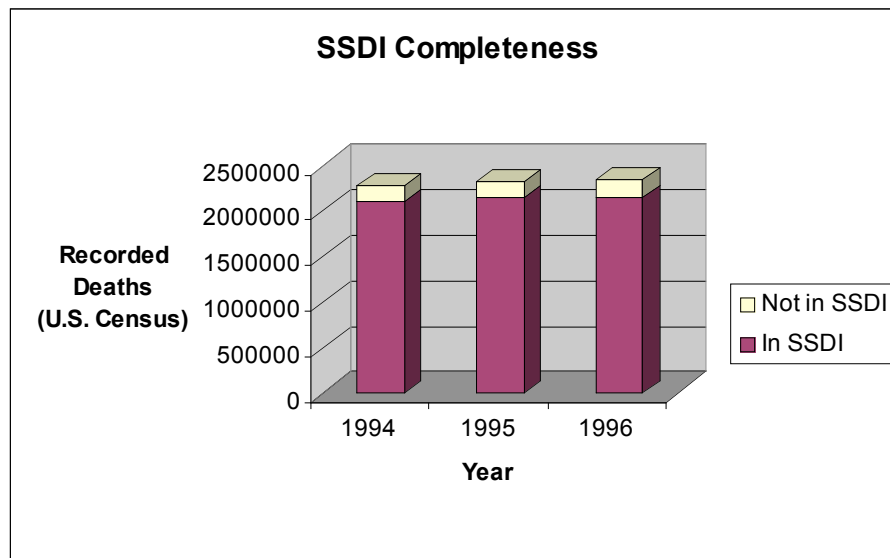
Figure 15: SSDI Completeness, 1994-1996

However, when we analyzed the data that we downloaded from RootsWeb, we noticed that certain years seemed incomplete. As seen in Figure 16, the number of Chicago SSDI records dropped significantly for 1990 and 1991, and probably 1989 as well. As the Ancestry.com interface reports almost exactly the same number of records, this is likely a problem with the SSA's DMF for this region and time.
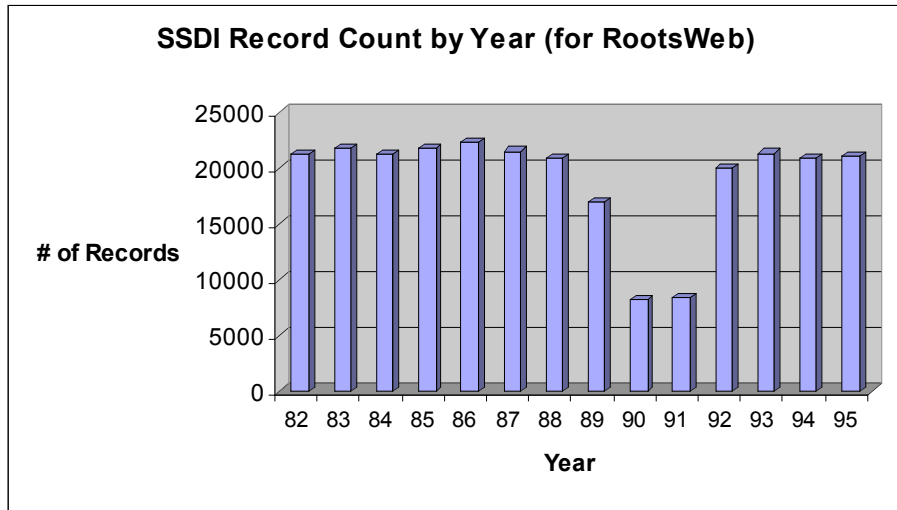
SSDI Record Count by Year (for RootsWeb)

Figure 16: SSDI Record Count, 1982-1995

To address our concern about the possibility of age skew, we compared the nationwide deaths by age range distribution for 1995 with the SSDI deaths by age range distribution for 1995. The results are shown in Figures 17 and 18. The SSDI is slightly underreporting deaths for young victims, but otherwise it closely matches the national distribution. One possible explanation for this is that many funeral directors will report deaths to the SSA as part of their services, counteracting the natural tendency of family members to not report young deaths in which no benefits would have been paid. As we saw in the statistics section for the Homicide data set, most homicide victims are young, which could mean that some of our eventual matches were false matches. (A record might match only a single row in the SSDI because similar rows were missing.) Keeping this potential problem in mind, we should be able to tailor our validation effort to the matches of which we are least sure.
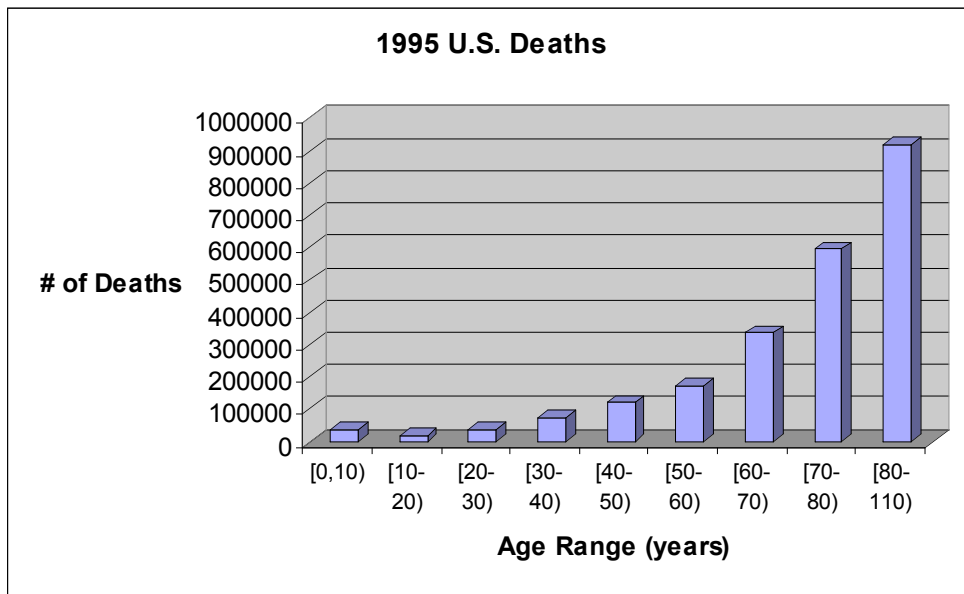
1995 U.S. Deaths

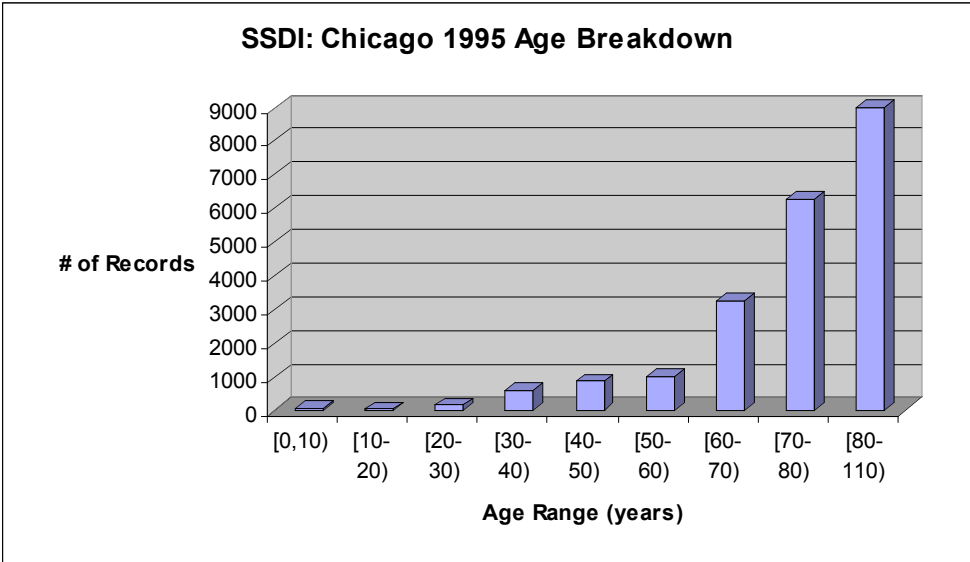**Figure 17: U.S. Deaths by Age Range, 1995**

Figure 18: Chicago Deaths in SSDI by Age Range, 1995

Joining the Databases

Joining the Chicago Homicide data set with the SSDI was attempted in four different ways with varying success. Initially, we tried to use geographic "hints" in the Chicago data to improve our matching, but this actually negatively impacted our matching. Our initial attempts also suffered from a mistake made in calculating the birth year of the victim. Our most successful method correctly matched birth years, and also used a third data set that mapped first names to gender. The four methods are described in detail in the sections below.

Initial Approach

Our initial approach at joining the Chicago Homicide data set with the SSDI was to look for instances where we could uniquely map {death year, death month, death day, victim's age, victim's location} tuples across both databases. However, we soon discovered that the fine-grained location information present in both databases was in incompatible formats. The SSDI included the state, county, and zip code of the last known residence while the Homicide data set included the police district number and the census tract number in which the murder occurred as well as information on whether the homicide occurred at the home of the victim or not. In addition, the entire homicide data set includes an implicit geographic identifier of state=IL, county=Cook, city=Chicago.

We began by restricting our analysis to only include individuals who died at home. These victims accounted for about 30% of the total victims. By doing so, we could guarantee that the location of the murder (which the data set told us) was the same location as the victim's last residence. We then acquired a mapping between census tract numbers and zip codes in Cook County from publicly available Census Bureau data (http://plue.sedac.ciesin.org/plue/geocorr/). This mapping is not without its faults; approximately 5% of the census tracts listed correspond to more than one zip code. These one to many mappings were eliminated. Finally, we used standard relational database management system technologies and methodologies to join the SSDI data along with the Homicide data, using our census tract to zip code mapping as an intermediary. The Structured Query Language (SQL) query that performs this joining is described in Appendix B as query 1.

The results were less than stellar. Out of 23,000 victims, only 10,000 had enough information to attempt a reidentification (only those who died after 1982). Furthermore, only about 3,000 of those actually died at home, allowing us to perform geographical linking. Finally, only 30 out of 3,000 were unique matches.

Our analysis suggests that our geographical mapping was flawed. In particular, we later discovered that geographical linking using zip codes is contraindicated, especially when looking at data that ranges over several decades. Zip codes were never designed to be used for geographic linking and suffer from a number of defects when used in this way. In particular, zip codes change quite frequently as they have no connection whatsoever to physical coordinates; they are merely mail routing designations, and as such are expected to change as delivery technology and city demographics evolve. In addition, unlike census tract numbers, they contain no versioning information; the zip code 02215 could represent a very different area in 1975 than it did in 1995 and there is no easy way to determine that.

Revised Approach

After seeing the problems that resulted from our attempts at fine-grained geographical matching, we attempted to reidentify individuals without fine-grained geographical matching. In particular, we removed all geographic restrictions on matches except the one that victims be residents of Chicago, Cook County, IL. This resulted in about 1,000 unique reidentifications out of 3,000 candidates. If we are willing to accept a higher error rate and presume that everyone in our sample was a Chicago resident, then we get approximately 7,600 unique matches out of about 10,000. The query that performs this matching is described in Appendix B as query 2.

However, upon further analysis, we discovered that our queries exhibited a subtle logic flaw. We initially assumed that we could calculate the birth year of a victim by subtracting their age from their death year. This calculation will yield the correct result approximately 50% of the time; it will be off by one year in the remainder of cases. We discovered that it is impossible to unambiguously calculate a birth year using only a death year and an age; all such cases have two possible birth years corresponding to them. This does not pose a significant problem for our reidentification experiments though, since our control data set indicates the precise birth date. Using this information, in most cases, we can resolve the ambiguity successfully. This modification increases the complexity of the query dramatically. The query itself is presented in Appendix B as query 3. This query provides 7,800 matches out of a total of 11,000 candidates.

Finally, we attempted to increase our matching rate even further by exploiting gender information provided by the Homicide data set. Unfortunately, the SSDI does not include a gender field. It does include a first and last name field. We used a public database from the Census Bureau containing rankings of the most popular first names in the United States in order to infer a gender for each SSDI record based on the first name listed. Before doing so, we had to strip out the names that were common to both genders. This extra information allowed us to resolve additional ambiguous matches, yielding 8,200 matches out of 11,000 candidates. The query that does this is listed in Appendix B as query 4.

Technical Specifications

In this chapter, we continue the technical explanation of our reidentification effort. We focus on the tools we used and why we chose them, our attempts to validate our identifications, and the effectiveness of two anonymizing techniques.

Tools

In order to conduct our reidentification experiments, we relied on a variety of tools. Our selection of what tools to use was constrained by a variety of requirements, some technical and some political. Because we had no real budget, we could only use freely available tools or tools we already own. Likewise, we could only afford relatively modest hardware on which to run our experiments, which meant that whatever tools we selected had to be relatively efficient. We needed to manipulate large amounts of data (often distributed using the SAS language) coming from many disparate sources (i.e., census bureau, geographical location info, homicide data) before actually performing the matching. The matching process required combining separate data sources based on a variety of common key matching strategies. We needed the ability to quickly change these strategies as we explored different matching techniques and incorporated new databases. Finally, because we had relatively little time in which to work, we needed to use tools that were either easy to learn or which we were already familiar with.

Based on these criteria, we chose to build our reidentification system using a relational database management system. The RDBMS approach gave us the flexibility we needed while at the same time allowing for reasonable performance and reduced development time. By taking advantage of the declarative semantics of the structured query language, we were able to leverage both our past experience in manipulating large databases of information and a time proven paradigm for using relationships to exploit patterns in large data sets. We further settled on the PostgreSQL relational database running on the Linux operating system. In addition, we developed a number of programs using the Python language to parse, clean, and load the data into the RDBMS. All of the tools mentioned so far were free.

We duplicated our data on a machine running Windows 2000 and another RDBMS, Microsoft's SQL Server 7.0. This allowed multiple people to work with the data, and provided redundancy in case of a catastrophic failure of the primary Linux system. We used Microsoft Excel 2000 to simplify data importing and exporting. We used Microsoft Visual FoxPro 6.0 to view the dBase IV formatted records of some deidentified data sets. Finally, ActivePerl build 618 fulfilled miscellaneous scripting needs.

Validation of Matches

After completing our reidentification experiments, we attempted to verify the efficacy and correctness of our reidentification techniques. This entailed comparing information in our reidentifications with publicly available information to ensure that the correct records were matched. In order to perform a complete verification, we would need an exhaustive register that listed all deaths in Chicago. While some states do make their death indices available online to the public (Texas and California for example), Illinois is not one of them. We are unable to locate any other authoritative death indices that could be used to verify our reidentification results.

If verification against an exhaustive registry is not possible, spot checks against a sparse registry might be effective. At the very least, they would give some information regarding the

reliability of our reidentification attempts. We are currently attempting to spot-check our results using newspaper stories and obituaries.

Anonymizing the Chicago Homicide Data Set

Sweeney describes several methods of anonymizing a data set in her seminal thesis[13]. As we did not have the time to test Sweeney's programs (see Suggestions for Further Work section), we tested three standard methods of anonymization.

1. Generalizing the victim age field to an interval of five years.
2. Generalizing the victim age field to an interval of ten years.
3. Removing identifiers as required for medical data by 45 CFR §164.514. (These identifiers are discussed in the Medical Protections section below.)

Figure 19 shows the results of our testing. 93.5% of Chicago homicide victims are uniquely identified by the {death year, death month, death date, gender, age} tuple. The first and second anonymization methods did not greatly reduce this uniqueness. (The tuples were 80.3% and 68.8% unique respectively.) The third anonymization method entailed stripping the death month and death date, and lumping all ages greater than or equal to 90 together. This method makes the resulting data only 4.7% unique.

The third method, removing the identifiers now forbidden for medical data, is the most effective anonymizing measure. However, this anonymity comes at a price; the resulting data cannot be used to analyze monthly homicide trends, which some researchers may wish to do.
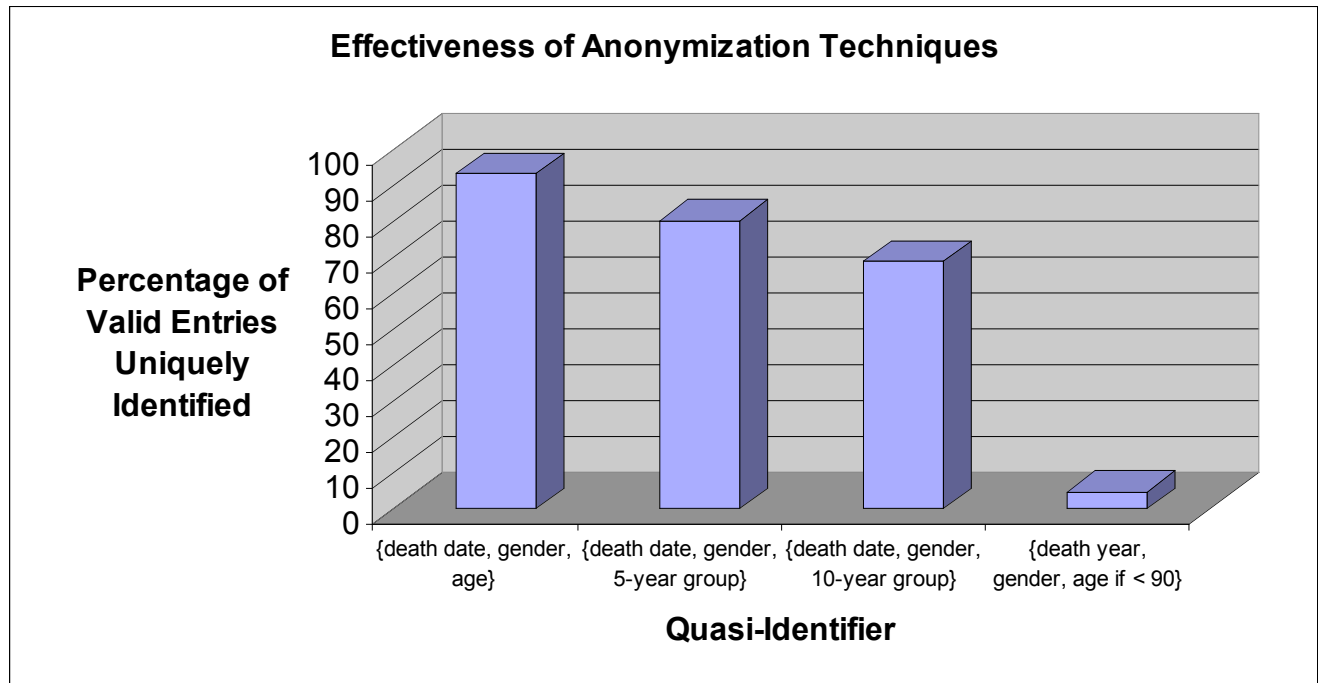

Figure 19: Effectiveness of Anonymization Techniques

---

[13] Sweeney, pg 60

Other Deidentified Data Sets

Before choosing a deidentified data set to focus on, we scoured the Internet for deidentified data from any source. We found several data sets of interest at the National Archive of Criminal Justice Data (NACJD) and from Investigative Reporters and Editors, Inc. (IRE)

The NACJD provides free downloadable access to hundreds of criminal justice data sets and analyses. It is one part of the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. The data sets it provides are culled from many sources including the federal and state governments.

In addition to the Chicago Homicide data set, the samples from the Robberies data set and the Arkansas Juvenile Courts data set were obtained on the NAJCD website. Data sets are provided in a variable-length-field format with SAS and SPSS codebooks.

IRE is a nonprofit organization that provides data and training to investigative journalists. They only sell their databases to journalists, journalism educators, and journalism students, but they make 100 record samples available for download. The government-sponsored data sets they provide are generally available directly from the respective agencies, but IRE standardizes the data format, and sells at or below the cost to them.

Data price is determined by the market size of the purchasing organization. The prices listed in the tables below are what students, freelance journalists, and periodicals with circulation below 50,000 would be charged.

The samples from the AIDS Patient data set and the Malpractice data set were obtained on the IRE website. Data sets are provided in the dBase IV format.

AIDS Patients

The AIDS Patient data set contains information about 688,200 individuals who have been diagnosed with AIDS since 1981. The information was originally collected by state and local health departments and was then collated by the CDC.

The information has been rigorously deidentified. Fields that might be useful for reidentification include the age group of the patient at the time of diagnosis, the month of diagnosis, the gender and race of the patient, whether the patient is currently alive, and the region of residence at the time of diagnosis. The region code corresponds to an area containing at least 500,000 people. AIDS patients who reside in less dense areas are simplify located as Northeast, South, Midwest, etc.

There are several fields that could be embarrassing to individuals who were reidentified, including whether or not the patient had sex with a bisexual man, whether or not the patient had sex with an injecting drug user, and so forth. However, we believe that this data has been deidentified so thoroughly that reidentification would be very difficult. Assuming that the percentage of the population diagnosed with AIDS is small enough, this database could possibly be joined with the Outpatient data set discussed in the next section. The resulting matches could them be joined with a control data set, though due to the vagueness of the region code, we doubt this would be a fruitful exercise.

Source:	Centers for Disease Control & Prevention, Division of HIV/AIDS Prevention
Size:	23.6 MB
Dates Covered:	1981 – 1998
Record Count:	688,200
Covers:	Entire United States
Cost:	$25

Figure 20: AIDS Patient Data Set at a Glance

Aids98

Age	Gender	Race	Categ	Dxdate	Repdate	Death	Exposure	Multrisk	Birth	Sexbi	Sexiv	Sexother	Sexhiv	Adjwgt	Msa
5	1	1	1	198810	198903	1	01	0	1	0	9	9	9	1.000	3360
6	1	2	7	199605	199608	1	01	0	1	0	0	0	0	1.041	1600
7	3	4	1	199704	199707	0	08	2	1	9	9	9	9	1.101	4480
5	1	2	4	199602	199603	0	01	0	1	0	0	0	0	1.033	1600
5	3	2	7	199701	199706	0	08	2	1	9	9	9	9	1.106	0003
6	1	3	1	199610	199703	0	01	0	1	9	9	9	9	1.062	5600
4	1	2	1	199505	199506	1	01	0	1	9	9	9	9	1.007	5080
6	3	2	3	199307	199308	1	08	2	1	9	9	9	9	1.000	0520
5	3	2	7	199501	199503	0	02	1	1	0	1	0	1	1.004	0003
5	1	1	7	199204	199303	1	01	0	1	9	9	9	9	1.000	7040
8	1	1	7	199810	199812	0	01	0	1	0	9	9	9	2.319	5960
5	4	2	7	199603	199605	0	02	0	1	9	9	9	9	1.016	0003
5	3	1	3	199401	199410	1	02	0	1	0	9	0	0	1.011	1600
6	2	2	1	199107	199110	1	01	0	2	9	9	9	9	1.000	5600
6	1	1	3	199505	199607	0	01	0	1	9	9	9	9	1.010	4520

Figure 21: Sample of CDC AIDS Patient Database

Outpatient Data

The information has been poorly deidentified. Latanya Sweeney has had remarkable reidentification success with this data set, especially when focusing on specific groups, like children with neuroblastoma. Fields that are useful for reidentification include the age, gender, marital status, and race of the patient, a region code and more.

The primary field that could be embarrassing or damaging to individuals who were reidentified is the diagnosis field. The diagnosis field contains an extremely detailed code for the patient's condition; the code could be used to reidentify women who had had abortions, or those infected with HIV.

Need more info on Latanya's control data set

Source:	National Center for Health Statistics (NCHS)
Size:	Huge
Dates Covered:	1965 - present
Record Count:	Huge
Covers:	Entire United States
Cost:	Varies, depending on provider and coverage

Figure 22: Outpatient Data Set at a Glance

```
9713002120310003l321    4813 V30017706-7671-               0319
9713002120310002l321    4815 V3000                          0319
9713002220410002l321    5583 V3000                          0419
9713002120410002l321    5583 V3000                          0419
9713001120410002l321    5586 V3000                          0419
9713001120510001l321    5663 V3000                          0519
9713002220610002l321    5663 V3000                          0619
9713001220710002l321    5813 V3000                          0719
9713002120710001l321    5815 V3000                          0719
9713001220710003l321    5816 V3000765107746-7793-           0719
9713002120810002l321    48663V31017706-                     0819
9713001120810001l321    4865 V30007671-                     0819
9713001120810001l321    4863 V3000                          0819
9713002220810002l321    4863 V3000                          0819
9713002120910003l321    4223 V3001                          0919
9713001520910002l321    4223 V3001                          0919
9713001120910002l321    4226 V30007731-                     0919
9713001121010003l321    4376 V3001                          1019
9713002121010002l321    4373 V3000                          1019
9713001121110001l321    4473 V3000                          1119
9713002121110003l321    4476 V300176408                     1119
9713002321110001l321    4473 V3000                          1119
```

Figure 23: Sample of Newborn Records in NCHS Outpatient Database

Malpractice

The Malpractice data set contains 227,541 records of medical malpractice suits filed or adverse action taken against individual practitioners. We find this data set particularly interesting because investigative reporters from the New York Daily News were able to reidentify individuals in it using court records and other data sets. The published story details their reidentification method in depth. It is interesting to note that under federal privacy laws only hospitals and a limited number of people in the health care field are allowed access to the raw (containing names) data. However, legislation to remove the restrictions has been proposed in response to the exposé.

The information in the publicly available data set has been deidentified reasonably well. Fields that might be useful for reidentification include a random practitioner ID that allows linking within the data set, an age group (in 10 year units), the work state, the home state, the field of license, and the decade of graduation from medical school.

There are several fields that could be embarrassing to individuals who were reidentified, including a code specifying the type of malpractice (e.g. unnecessary tests of surgery or wrong body part), the amount of payment, and any other adverse actions, such as revocation of license or denial of professional society membership.

While the data set might not be reidentifiable using only online data sources, the efforts of the investigative journalists show the possibilities inherent in reidentification.

Source:	U.S. Department of Health and Human Services
Size:	37 MB
Dates Covered:	1 Sep 1990 – 31 Dec 1999
Record Count:	227,541
Covers:	Entire United States
Cost:	state slice, $20; entire U.S., $55

Figure 24: Malpractice Data Set at a Glance

Recno	Workstat	Homestat	Licnstat	Licnfeld	Agegroup	Gradgrup	Malcode1	Malyear1	Payment	Practnum
14	CA	CA	CA	10	30	1980	240	1989	$12,500	3138
15	SD	SD	SD	10	40	1960	10	1988	$845,000	35191
16	LA		LA	10	50	1950	290	1985	$17,500	15272
17	AR		AR	10	30	1980	280	1991	$92,500	960
18	GA		GA	10	40	1960	280	1990	$47,500	10159
19	MI	MI	MI	20	40	1960	590	1988	$42,500	43198
20			TX	10		1940	610	1974	$105,000	37355
21	TX		TX	10	40	1960	250	1991	$8,750	36440
22	TX		TX	10	60	1950	665	1990	$195,000	58117
23	PA		PA	10	30	1970	620	1982	$2,500	70752
24	FL	FL	NM	10	30	1980	10	1993	$32,500	70799
25	FL		FL	10	30	1970	690	1989	$145,000	71371
26	MI		MI	10	40	1970	190	1992	$145,000	70234
27	FL	FL	FL	10	30	1980	710	1990	$17,500	66115
28	CA	FL	FL	10	30	1980	250	1991	$6,250	72802
29	CA		CA	10	60	1950	250	1993	$72,500	70415
30	TX	TX	TX	110	50	1960	120	1992	$22,500	71080

Figure 25: Sample of Department of Health and Human Services Malpractice Database

Chicago Robberies

The Chicago Robberies data set contains information about 7,216 robbery victims. It is split into several parts based upon how injured the victim was. (Not injured, injured, killed)

The information has been reasonably deidentified. Fields that might be useful for reidentification include the age group of the victim, the gender, race, and marital status of the victim, the employment status of the victim, and the victim's district of residence.

There are several fields that could be embarrassing to individuals who were reidentified, including whether or not the victim was dealing drugs, whether or not the victim was in a gang, and a code describing the relationship between the victim and the offender, similar to that found in the homicide data set. However, because the age range is so vague (infant, young adult, etc.) we do not believe this data could be easily reidentified. Reidentification would also be difficult because of the relatively small size of the data set.

Source:	Inter-university Consortium for Political and Social Research (ICPSR)
Size:	759 KB
Dates Covered:	1982-1983
Record Count:	7,216
Covers:	Chicago
Cost:	Free

Figure 26: Chicago Robberies Data Set at a Glance

Figure 27: Sample of ICPSR Chicago Robberies Database

Juvenile Court Records

The Arkansas Juvenile Court Records data set contains information about 55,467 juvenile offenses committed between 1991 and 1994. Juvenile Court Records with virtually identical information are available for Missouri for the years 1984-1987 and 1994, and for Nebraska for the years 1975-1994. The raw records included the juvenile's name, attorney, etc.

The information has been very poorly deidentified. Fields that might be useful for reidentification include the year and month of birth of the juvenile, the gender and race of the juvenile, and the county of jurisdiction. Some data sets also include the exact date of birth and the census track of residence, though that information is not present in all data sets, and it has been stripped from some others to inhibit reidentification.

The primary field that could be embarrassing is the offense field. This is a code specifying the type of crime committed. Possible values for the Arkansas code include everything from "capital murder" and "rape" to "unlawful packaging of strawberries" and "use of X-ray shoe-fitting machines".

In the figure below, the fourth column of text is the race code, the fifth column is gender, and the six through eleventh columns are the date of birth in YY|MM|DD format.

We are confident that individuals can be reidentified in these data sets. Latanya Sweeney found that 87% of the population in the United States is likely unique based only on {5-digit ZIP, gender, date of birth}. We believe that a similar or higher percentage or juveniles could be reidentified given this publicly available data.

Source:	Arkansas Administrative Office of the Courts
Size:	7.1 MB
Dates Covered:	1991 – 1994
Record Count:	55,467
Covers:	Arkansas (other states publish this info as well)
Cost:	Free

Figure 28: Juvenile Court Records Data Set at a Glance

```
01D  F77  323RDN94 2 4 0 0 0 98998999898    989899899988    98989989988  9894 3 4    0   0    OFN    0 B  0
01DWF79  4 9RDN94 2 4 0 0 0 98998999898    989899899988    98989989988  9894 3 4    0   0    OFN    0 B  0
01DWF81  7 9RDN94 2 4 0 0 0 98998999898    989899899988    98989989988  9894 3 4    0   0    OFN    0 B  0
01DWM80 523RDN94 2 4 0 0 0 98998999898    989899899988    98989989988  9894 3 4    0   0    OFN    0 B  0
01DWF8212 1RDN94 2 4 0 0 0 98998999898    989899899988    98989989988  9894 3 4    0   0    OFN    0 B  0
01DWM78  2 9RDN94 2 4 0 0 0 98998999898    989899899988    98989989988  9894 3 4    0   0    OFN    0 B  0
01DWM76 523RDN94 2 2 0 0 0 98998999898    989899899988    98989989988  9894 3 4    0   0    OFN    0 B  0
01D  F7812 5DJD931020 0 0 0  5 36 10398MA 19899899988    98989989988  9894 2 4    0   0    ODM    0 N  0
01D  M77 317DJD931026 0 0 0  5 36 10398MA 19899899988    98989989988  98941117    0   0    OOD    0 N  0
01D  M77 725DJD9311 3 0 0 0  5 38 20498MA 1 5 13 3018FD 1 5 132078MC 194 1 5    0 35    OFD    0  Y 6
01D  M93 519DDN931210 0 0 0 98998999898    989899899988    98989989988  9894 110    0   0    OFN    0 B  0
01D  F89 512DDN931221 0 0 0 98998999898    09899899988    98989989988  98941018    0   0    OOD    0 N  0
01D  F80 327DDN931221 0 0 0 98998999898    09899899988    98989989988  9894 314    0   0    OFN    0 N  0
01D  F81 823DDN931221 0 0 0 98998999898    09899899988    98989989988  9894 314    0   0    OFN    0 N  0
01DWM76 530TJD94 1 5 0 0 0  5 36 10698FC 19899899988    98989989988  9894 1 5    0   0    OFDD   0 N  0
01DBM78 522DJD94 111 0 0 0  5 13 20498FD 1 5 71 2078MC 1 5 132038MA 294 112    0   0    OFDD   0YN  0
01D  F80  6 4DFS94 125 0 0 0 98998999898    989899899988    98989989988  9894 228    0   0    OFF    0 N  0
01D  F81  7 3DFS94 125 0 0 0 98998999898    989899899988    98989989988  9894 228    0   0    OFF    0 N  0
01D  9-19999DDN94 3 1 0 0 0 98998999898    989899899988    98989989988  9894 314    0   0    OOD    0 N  0
01D  9-19999DDN94 3 1 0 0 0 98998999898    989899899988    98989989988  9894 314    0   0    OOD    0 N  0
01D  F80 721DJD94 322 0 0 0  5 36 10398FB 19899899988    98989989988  9894 427    0   0    OFD    0 BY12
01D  F801127DJD94 322 0 0 0  5 36 10398FB 19899899988    98989989988  9894 427    0   0    OFD    0 BY12
01D  F77 618DJD94 322 0 0 0  5 36 10398FB 19899899988    98989989988  9894 427    0   0    OFD    0 BY12
01DWM80 223DJD94 322 0 0 0  5 36 10398FB 19899899988    98989989988  9894 427    0   0    OFDO   0 BY12
01D  F761219DJD94 330 0 0 0  5 71 20798MC 1 5 71 2128MC 1989989988  9894 429 200   0    OFD    0 BY12
01DWM77 725TJD94 331 0 0 0  5 38 20398FC 2 5 39 2018FB 1 5  32018FC 194 331    0   0    OFDD   0 B  0
```

Figure 29: Sample of Arkansas Administrative Office of the Courts Juvenile Database

Other Control Data Sets

The SSDI was the only control data set we could find that was a suitable match for the Chicago Homicide data set. Nonetheless, other control data sets are available on the Internet. This chapter describes two data sets that could be useful as control data sets in reidentification experiments, voting records and vital records.

Voting Records

Online voting records are not universally available. Most online voting records can only be found for individual counties. The voting records shown in the figure below are for the Dallas County, Texas area [IORP]. Voting records are useful when the deidentified information concerns living persons and the reidentifier would like contact information for those persons.

In the Dallas County records, there are fields for the voter's name, address, date of registration, and certification number. In addition, for each voter, information is kept on the dates he voted, what the election was for (e.g. Governor, Presidential), the party affiliation he declared, and the manner of voting (e.g. In person, Early). Other counties usually have these fields in their records too, though some also include the voter's gender and birth date. As seen in the first three records in the figure below, voting records are sometimes incomplete, containing a first initial rather than a first name.

```
Name              Address                                    Registered   Certification
A P SMITH         05102 SACHSE, Sachise  75048               10/06/91     02010756
A D SMITH         01618 PRESIDIO, Dallas  75216              11/01/80     01133182
A C SMITH         06322 MOONGLOW, Dallas  75241              10/01/76     00699232
AARON M SMITH     05323 N MACARTHUR, Irving   75038          01/08/98     02704875
AARON I SMITH     01907 MEADOW DALE, Irving   75060          07/27/98     02744886
AARON T SMITH     05541 NAKOMA, Dallas  75209                08/13/93     02260436
AARON L SMITH     16310 RED CEDAR TRAIL, Dallas   75248      07/13/94     02322925
AARON R SMITH     08927 SAN BENITO WAY, Dallas   75218       02/09/92     02033167
AARON D SMITH     02729 E HWY 80 259, Mesquite   75150       03/05/98     02714117
AARON K SMITH     00740 PLEASANT HILLS, Dallas   75217       01/02/80     00927153
AARON SMITH       00985 GREEN CASTLE, Dallas   75232         10/01/80     01063141
AARON A SMITH     07849 WOODSHIRE, Dallas   75232            02/12/96     02503582
ABNER C SMITH     01002 S BRITAIN, Irving   75060            03/01/76     00087297
ADA J SMITH       02500 W 7TH, Irving   75060                11/03/92     02163992
ADA R SMITH       00222 S BRISCOE, Dallas   75211            03/01/76     00268572
ADA SMITH         00915 ST ANDREWS, Desoto  75115            08/01/77     00768765
```

Figure 30: Sample of Dallas County Voting Records

Birth/Death/Marriage/Divorce Records

Several states have free online indices to their birth certificates, death certificates, marriage records, and divorce records. While this section will focus on the records available for the state of Texas [TDH], similar records are available for Alabama, California, Kentucky, Vermont, and other states, sometimes for free, sometimes for a fee.

The fields in the online indices vary from state to state. For example, California's online death index includes the name of the deceased, birth date, birthplace, death date, death place, gender, mother's maiden name, and occasionally social security number. In contrast, Alabama's online death index only includes the name of the deceased, county of death, date of death, and the state certificate number.

An excerpt from the 1999 Texas death index is shown in the figure below. The fields are last name, first name, middle name or initial, death date, county where death occurred, and gender. This is the shortened format used between 1976 and 1999. For the prior 12 years, additional fields such as social security number, marital status, and spouse's name were recorded in the index.

Texas also publishes its marriage and divorce records. The marriage records contain the name and age of each partner, along with the marriage date and county. The divorce records contain similar name, age, and marriage date information, and also include the divorce date, the number of children under 18, and the county where divorce occurred.

The Texas records are especially suited for reidentification experiments as they are downloadable in bulk by year, while many sites only provide a searchable interface, and not direct access to the data. The Texas records are in tab-delimited format, further simplifying their importation into a RDBMS.

The non-availability of an online index to the Texas birth records is interesting to note because of another Internet privacy concern. Texas state law only allows the publication of birth records if they cannot be used to identify information about an adoption. The Texas birth index was available online until the Texas Bureau of Vital Statistics identified a case in which the index may have assisted in the identification of an adoption. It has since been taken down from the official site pending review, though due to its online sojourn it is now mirrored in other online locations.

Source:	Texas Department of Health, Bureau of Vital Statistics
Size:	60.8 MB
Dates Covered:	1976-1999 summary information, 1964-1975 general information
Record Count:	About 4.4 million
Covers:	Texas
Cost:	Free

Figure 31: Texas Death Record Index at a Glance

Aaron	Dora	Pruett	19990217	ATASCOSA	FEMALE
Aaron	Mamie	Laverne	19990219	TARRANT	FEMALE
Aaron	Dorothy		19990705	BELL	FEMALE
Aaron	Howard	L	19990706	HARRIS	MALE
Aaron	Sandra	Gayle	19990729	DALLAS	FEMALE
Aaron	Chester	Laray	19990813	TARRANT	MALE
Aaron	Bobby	Houston	19990826	TARRANT	MALE
Aaron	Essie	Louise	19990914	HARRIS	FEMALE
Aaron	Billy	Imogene	19991005	TARRANT	FEMALE
Aaron	Wilma	Simmy	19991012	HARRIS	FEMALE
Aaron	Carl	Charles	19991129	GREGG	MALE

Figure 32: Sample of Texas Death Record Index for 1999

We believe that these indices could be used as control data sets for reidentification. The possibility of privacy abuses is apparent from the misuse of the Texas birth index by itself. Potential privacy threats will increase as additional states make their vital record indices and other databases available online. The next chapter presents an analysis of the existing and

proposed legal ramifications of reidentification, which may curb some of the potential abuses we have discussed.

Legal Analysis

Are we breaking the law?

Clearly, the first and most important question we must ask ourselves when beginning a project such as this is, "are we breaking the law?" Any project involving private, personal information is naturally in dangerous territory. The core principle of the Privacy Act is that private information should remain private, and so, in theory, our project should be quite unlawful.[14]

From a philosophical point of view, however, we are simply using statistics for research, which is why they are published in the first place. In fact, in so far as it is our right to uncover problems with current government procedure, is quite possible that we could defend our actions in court with the First Amendment. The bottom line however, is that current legislation is rather ambiguous about our situation.

What if we tried to use the information?

Current privacy law deals primarily with information that has been entrusted to another party. The recipient of the personal information is under an obligation not to disclose it to anyone else without notice and (usually) permission. For example, recent privacy legislation prevents your local DMV from selling your address to marketers without your permission.[15]

Reidentification, however, does not quite fall within the scope of such laws because we have discovered the information for ourselves; no one has entrusted it to us, so we have no one's trust to break. In fact, as researchers, we are entitled to use the results of our project for publication, and, again, are probably protected by the First Amendment. Considered another way, our situation may be somewhat analogous to that of a reporter who uncovers personal information and then sells it to tabloids, a legal occupation. As we will see, the specific restrictions on how we can use reidentified information depend on what sort of information it is, and how we got it.

As a company, would this be breaking the law?

Companies fall under roughly the same restraints as private citizens in this regard, with the exception of credit bureaus, whose complicated restrictions we will not address. Companies are not prohibited from combining public databases for reidentification purposes, nor are they necessarily prohibited from using or selling the results as they see fit; again, the restrictions depend on what the information is, and where it came from. There is, however, a clear potential here for highly specific (and perhaps invasive) direct marketing.

As mentioned earlier in the paper, reidentification could be very easy for companies that maintain large customer databases. Reidentifying customers to find out more about them is not unlawful. More importantly, this scenario would be very difficult to police, because any company downloading databases could be doing so for honest statistical analysis. Once the information is stored on a private computer, there is no way to track what it is being used for; law enforcement would have to literally catch the company in the act.

---

[14] Privacy Act of 1974. 5 USC §552, 552a, 552b.
[15] Driver's Privacy Protection Act. 18 USC §2721-25.

Companies are prohibited, however, from selling reidentified data that used their own customer databases as controls. Because the data is derived, in part, from customer-entrusted private information, company use of the data is restricted in the same way as the original information.

As the government, would this be breaking the law?

On this point, at least, the public can feel secure in their privacy. The Privacy Act specifically prohibits agents of government from combining databases with each other. Immigration, for example, is not allowed to ask the IRS whether or not you've paid your taxes before renewing your green card.[16]

Looking at US Laws

Medical. Privacy of medical information is a source of great public concern. Protections exist to prevent misuse of the information by hospitals, HMO's, and Insurance companies.[17] Medical information tends to be very personally sensitive, and this is probably why it has the most legislation protecting it. In addition, a new regulation designed to hinder reidentification establishes standards for deidentification of medical information.[18]

Criminal. Criminal information, as we shall see, is much less well-protected then medical information. While provisions exist in United States Code[19] and in the Code of Federal Regulations[20] to regulate its distribution and use, standards for deidentification are not clearly defined. In addition, as we have shown in our project, criminal databases are published in a format that allows reidentification of crime victims.

Other Information. The Privacy Act contains some information protections, however it is mostly concerned with preventing agents of the government from combining databases with each other. Special protections do exist for some other private information; the Fair Credit Reporting Act, for example, protects financial information.[21] For the most part, however, the only information with legislation to protect it is information that has been sufficiently misused to provoke a public response.

Privacy Act vs. FOIA conflict

The fundamental difficulty of legislating against reidentification stems from the conflict between the Privacy Act and the Freedom of Information Act. When does the public right of access to information supersede the individual's right to privacy? While this issue has been a source of considerable debate in the courts, the specific issue of reidentification has been given small notice.

---

[16] Privacy Act of 1974. 5 USC §552, 552a, 552b.
[17] Hospital Insurance Accountability and Portability Act of 1996 (HIPAA). Public Law 104-191.
[18] Standards for Privacy of Individually Identifiable Health Information. 45 CFR §164.514(b).
[19] 42 USC §3789g, 3732, 10505.
[20] 28 CFR §22.; 28 CFR §513.
[21] Fair Credit Reporting Act. 15 USC §1681.

Southern Illinoisan vs. DPH

The case of the Southern Illinoisan vs. the Department of Public Health[22] is one of the only examples of reidentification coming up in court. In 1997, the Southern Illinoisan, publisher of a highly circulated regional newspaper, requested statistical information on incidence of neuroblastoma from the Department of Public Health under the Freedom of Information Act. The DPH denied the request, asserting that information was excepted in section 7(1)(b)(i) of FOIA, because its release would "constitute a clearly unwarranted invasion of personal privacy." The DPH argued that, as demonstrated by Sweeney's work, it was possible to reidentify individuals from the data, which included diagnosis date, cancer type, zip-code.

The DPH was ordered by a Circuit Court to release the documents, and DPH subsequently appealed to the District Court. In addition, the Circuit Court refused to admit as evidence an affidavit of Sweeney's, which it deemed conclusory. The District Court, upon reviewing the facts, disagreed with the Circuit Court's summary judgment, insisting that it consider Sweeney's affidavit, and remanded the case. In the opinion of the court, released March 28, 2001, Justice Chapman insisted that the lower court address the issue of whether "the information sought [will] reasonably tend to lead to the identity of any person whose condition or treatment is submitted to the Cancer Registry."

Privacy Act

The Privacy Act of 1974[23] was enacted in response to public concern over the growing number of government databases. The potential for government agencies to combine databases was clear, and Americans were afraid of big government intruding into citizens' private lives. The primary goals of the act were preventing government agencies from sharing databases with each other and giving rights of access and correction to individuals who were listed in databases, both government and private.

Criminal Protections

In addition to Privacy Act protections, criminal records and statistics are governed by a complicated mesh of laws. Provisions in United States Code Title 42 stipulate that criminal history information be kept up to date, and as complete as possible. This legislation stipulates the same rights of access and correction as the Privacy Act. The Bureau of Justice Statistics[24] is charged with maintaining crime databases and establishing standards for the collection, storage, and public distribution of the information. BJS is specifically charged to "provide information to … the general public on justice statistics."[25]

The regulations for distribution of criminal database information are somewhat complex, due largely to government funding of research specified in the Crime Control Act[26], the Juvenile Justice Act[27], and the Victims of Crime Act[28]. Research groups may apply for funding and data from government agencies such as BJS, and must enter into an

---

[22] The Southern Illinoisan v. the Department of Public Health. Appellate Court of Illinois
[23] Privacy Act of 1974. 5 USC §552, 552a, 552b.
[24] http://www.ojp.usdoj.gov/bjs/
[25] 42 USC §3789g(c)(10).
[26] Omnibus Crime Control and Safe Streets Act of 1968. Public Law 90-351.
[27] Juvenile Justice and Delinquency Prevention Act of 1984. 42 USC §5672.
[28] Victims of Crime Act of 1984. 42 USC §10601-§10608.

"Information Transfer Agreement" to do so.[29] This agreement, specified in the federal regulations, requires the recipient of the data to:

- Use the data only for research or statistical purposes
- Not reveal the data "to any person for any purpose except … in research findings (and/or databases) … on a need-to-know basis for research or statistical purposes…"[30]
- Store the data securely
- Design research projects to "preserve the anonymity of private persons to whom the information relates"[31]
- Not disseminate "information which can reasonably be expected to be identifiable to a private person"[32]
- Return or destroy the information after completion of the project

The website we obtained the Chicago Homicide database from is managed by the National Archive of Criminal Justice Data, an archive of the Inter-university Consortium for Political and Social Research.[33] NACJD maintains and provides public access to the criminal databases compiled by BJS. The organization is federally funded in accordance with the aforementioned regulations and is bound by the Information Transfer Agreement. Upon entering the online NACJD archives, a "data use restriction" agreement prompts the user to agree to the Information Transfer Agreement by entering an email address. It is possible to bypass this "data use restriction" agreement, however, by linking directly to the data pages, since the website does not employ any means to check if a user has passed through the disclaimer page.

The publication of the Chicago Homicide database on the Internet would seem to be in violation of the "need-to-know basis" clause in the CFR. The information is available to anyone, and the user is not even, necessarily, notified of the Information Transfer Agreement. However, the NACJD is fulfilling BJS's duty to provide justice statistic information to the general public, a requirement of federal law, which supersedes the CFR.

While the method of publication of the BJS databases may be deemed lawful, the format of the data raises more important questions. The CFR sections on criminal and justice statistics includes the following definition:

> 28 CFR §22.2 (e) Information identifiable to a private person means information which either-- (1) Is labelled by name or other personal identifiers, or (2) Can, by virtue of sample size or other factors, be reasonably interpreted as referring to a particular private person.[34]

Subsection (2) clearly refers to reidentifiable information. This information is specifically prohibited from publication by the CFR.[35] Since, as we have shown in our project, reidentification of individuals from NACJD data sets is quite possible, NACJD is in violation of its Information Transfer Agreement with BJS.

---

[29] 28 CFR §22.24 Information transfer agreement.
[30] 28 CFR §22.24(b).
[31] 28 CFR §22.24(f).
[32] 28 CFR §22.24(g).
[33] http://www.icpsr.umich.edu/NACJD/welcome.html
[34] 28 CFR §22.2 Definitions.
[35] 28 CFR §22.24(g).

If the Chicago homicide database has been designated by BJS as not individually identifiable, then it would be exempt from the Information Transfer Agreement. However, since the data is clearly identifiable, BJS would be in violation of federal regulations for releasing the information without the binding of the Information Transfer Agreement. In either case, it is clear that the release of this information is a violation of federal regulations.

NACJD seeks, with the "data use restriction" agreement on its web site, to bind any users who access the information by the Information Transfer Agreement specified in the CFR. While the legality of a click-through user agreement is questionable, the ability to bypass the agreement entirely certainly annuls it; you cannot be held responsible for an agreement you have never seen or signed.

Were we, however, to be held accountable for the Information Transfer Agreement, we would have the following use restrictions:

- "Research or statistical information identifiable to a private person may be used only for research or statistical purposes."[36]
- "Project plans will be designed to preserve anonymity of private persons to whom the information relates..."[37]
- "Project findings and reports prepared for dissemination will not contain [identifiable information]"[38]
- "Research or statistical information identifiable to a private person shall be immune from legal process and shall only be admitted as evidence... with the written consent of the individual to whom the data pertains."[39]

Although we are using the information purely for research, we have failed to "preserve anonymity of private persons," (that being the goal of our project), and so we would clearly be in violation of that clause. However, our intentions are not malicious, and we neither intend to publish the names of re-identified homicide victims, nor present our findings in court. As mentioned before, it is likely our actions are defensible under the first amendment, since our project goal is critiquing governmental procedure.

Medical Protections

As mentioned previously, medical information is the most carefully protected private information. The Hospital Insurance Accountability and Portability Act of 1996 (HIPAA)[40] is responsible for phenomenal new standards in patient privacy. In particular, the Department of Health and Human Services has just passed HIPAA compliance regulations designed to prevent reidentification.

The new CFR regulations, entitled "Standards for Privacy of Individually Identifiable Health Information,"[41] came into effect on April 14 of this year. The regulations give specific requirements for de-identifying medical information. In addition to requiring the removal of standard identifiers, such as name, address, and social security numbers, the regulations require the removal of the following:

---

[36] 28 CFR §22.21 Use of identifiable data.
[37] 28 CFR §22.24(f).
[38] 28 CFR §22.24(g).
[39] 28 CFR §22.28 Use of data identifiable to a private person for judicial, legislative or administrative purposes (a)
[40] Hospital Insurance Accountability and Portability Act of 1996 (HIPAA). Public Law 104-191.
[41] 45 CFR§162, §164.

45 CFR§164.514 (2)(i)(B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.

45 CFR§164.514 (2)(i)(C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

All publishers of medical information have until February 26, 2003 to implement these new standards.[42] These requirements for deidentification are very effective at anonymizing patients. A population size of 20,000 effectively kills all chances of reidentification by location. In addition, the restriction of all date information to year makes it impossible to do the sort of date-based matching that our project employed.

Proposed Legislation: Medical

Privacy is a hot issue right now in the United States, clearly evidence by the drastic increase in privacy legislation proposed in state and federal governments. Unfortunately, in the fervor of legislation, reidentification issues have been little addressed. Excluding the recently passed "Standards for Privacy of Individually Identifiable Health Information," only one other bill at the federal level specifically addresses identifiable information issues.

The Medical Information Protection and Research Enhancement Act of 2001 contains a clause permitting the "use [of] protected health information for the purpose of creating nonidentifiable health information."[43] "Protected health information" refers to any information relating to the health of an individual, or derived from their medical records. This law would seem to compliment "Standards for Privacy of Individually Identifiable Health Information," by making it easier to create public statistical databases.

The proposed law does, however, contain a clause that could be construed as a loop-hole in the HIPAA compliance regulations 45 CFR 160-164: "A person … may disclose [protected health information] to a health researcher if the research project has been approved by an institutional review board"

"Institutional review board" is not defined, and it seems possible that entities who cannot do reidentification because of the new database standards, could still get access to private information.

---

[42] The February 26, 2003 deadline excepts "small health care providers," who have until February 26, 2004 to implement the standards.
[43] Medical Information Protection and Research Enhancement Act of 2001. Proposed Bill H.R.1215.
http://thomas.loc.gov/

A German Reidentification Law

United States takes a sort of piece-meal approach to privacy legislation; if some type of personal information is causing privacy problems, a highly specific law will be drafted to deal with the problem. Other countries, however, take different approaches. Germany, for example, has recently passed a short but pointed reidentification law. In its entirety, the text of the law states:

> It is prohibited to match individual data from federal statistics or to combine such individual data with other information for establishing a reference to persons, enterprises, establishments or local units for other than the statistical purposes of this Law or of a legal provision ordering a federal statistics.[44]

German law also provides a penalty of up to one-year imprisonment or a fine for infringements.[45]

While a law this general would effectively protect against reidentification, it is unlikely it would be passed in the United States. As mentioned above, use of public information, including statistics, is quite clearly protected by the First Amendment. Legislation such as this would likely be deemed unconstitutional.

---

[44] Law on Statistics for Federal Purposes, Article 21
[45] Law on Statistics for Federal Purposes, Article 22

Legal Recommendations

New technology and the pervasiveness of the Internet and electronic data are turning the public eye toward privacy issues with increasing concern. As we have shown, current privacy legislation is insufficient to prevent reidentification of individuals from public databases. More complete legislation with better specifications is needed.

While the United States would have a difficult time passing reidentification laws similar to Germany's, legislation should be drafted to protect against misuse of information, regardless of the source. It should not matter whether private information being misused was entrusted to the abusive party, or discovered by them. We recommend that reidentified information be treated in the same way as entrusted information in this regard.

Additional protections of data are also needed. In particular, misuses of data should be better defined, and the victims given course of redress. Companies (like health insurance providers, and credit bureaus) should be prohibited from combining their own databases with publicly available ones to learn things about their customers; this should classify as improper use of entrusted personal information, and customers should be able to sue.

Federal standards should be adopted for de-identifying information. At a minimum, the standards introduced in Standards for Privacy of Individually Identifiable Health Information should be applied to all publicly release statistical databases, including the criminal statistics. Specifically, age and location aggregation should be required. If the information is sufficiently deidentified, cases like the Southern Illinoisan vs. Department of Health can be avoided. Determining release of information based on whether it "constitutes an unreasonable privacy invasion"[46] would no longer be an issue. Eliminating the possibility of reidentification is key to resolving the conflict of interests between FOIA and the Privacy Act.

In addition, it is clear that the Chicago homicide database and other similar databases violate federal regulations. Not only is the "data use restrictions" agreement insufficient to bind private parties from misusing the information, the reidentifiable format in which the data is published is unlawful. These databases should be removed from public access. They must be further deidentified before publication.

Technical Recommendations

Technical recommendations are easy to state but hard to implement. Before releasing any deidentified data, the holder of that data should consider the possibility of reidentification. Having a thorough knowledge of the available control data sets is especially useful when determining which fields in the deidentified data set will still be problematic.

As we saw in the section on Anonymizing the Chicago Homicide Data Set, there are simple and effective methods to limit reidentification. Specifically, removing the identifiers specified by 45 CFR§164.514 (2)(i)(B & C) seems to work. We recommend taking the Chicago Homicide data set offline until a thorough anonymization can be performed. We also strongly recommend taking the Juvenile data sets offline until a thorough anonymization can be performed, as there are grave privacy threats to living persons with information in those databases.

---

[46] The Southern Illinoisan v. the Department of Public Health. Appellate Court of Illinois

Suggestions for Further Work

Due to time limitations, we could not implement all of our ideas. Our suggestions for further reidentification work with the Chicago Homicide data set include:
- Find a better validation data set, and use it to compute our reidentification success rate. As described in the validation section, we cannot compute our success rate using online obituaries for the Chicago area. A better validation data set would be an index of death records for the area.
- Apply the datafly, μ-argus, and k-similar systems to the Chicago Homicide data set. Compare the systems by their execution speed, the completeness of anonymization, and the usefulness of the resulting data. Also compare the systems' performance to that of the more traditional anonymization techniques we describe in the anonymization section. Evaluate the actual reidentifications possible after these anonymizations, and not just the uniqueness of the quasi-identifiers.

In addition, we believe that many of the reidentification tasks we were faced with could be automated for use with other deidentified data sets.

Conclusion

The data explosion continues. The amount and availability of all types of personal information that is being collected is increasing at a tremendous rate as a result of continuous technological advancements. These advancements not only allow for the storage of more data, but also its ease of transfer. The motivations for data collection are numerous, ranging from medical to marketing and more. Data holders are being driven to share their data with others in order to spur more research, support ongoing research, or for any of a multitude of other reasons depending on the nature of the information they hold. However, much of the information is not only personal, but also private.

For this reason, data holders need to consider their subjects' privacy as they make their databases available or otherwise release their information. They are then left with a dilemma: should they make the data more anonymous, rendering the data less useful, or should they make the data as useful as possible, sacrificing more of the subjects' privacy? They need to find an optimal release of data – one that maximizes both usability and anonymity of the data. However, we, as data investigators, have found that databases are often made public under the false assumption that they have been rendered completely anonymous by the fact that they have been deidentified. Data holders, supported by statisticians' assurances, have come to believe that simply removing explicit identifiers from data makes them anonymous.

Research by Latanya Sweeney and others proves that deidentified data is not equivalent to anonymous data. Anonymous data cannot be manipulated to reidentify individuals, whereas deidentified data can be. Using quasi-identifiers one can easily uniquely identify a large number of the subjects of a deidentified data set. Reidentification makes use of these quasi-identifiers in linking the subjects to individuals named in a control data set.

Our experiment involving reidentification of victims in a homicide database demonstrates the relative ease through which one can reidentify data subjects. Sweeney has focused much of her research on medical data, causing medical workers and lawmakers to focus their attention on new standards for the release of patients' private medical information. Her research even forced the usually slow wheels of government bureaucracy to turn more quickly, allowing a new set of

federal regulations to be drafted and set in place. We took on the challenge of exploring reidentification in a whole other field by attempting to reidentify victims in a criminal database. We did this to not only bring attention to databases that contain criminal information but to illustrate that the anonymity problem extends to many other types of databases beside medical or criminal. The section on other deidentified data sets provided in this document points out other data sets that contain information that is as equally private as information in the homicide database we used in our experiment and that is equally subject to reidentification.

If reidentification only resulted in matching a small number of subjects to personally identifiable information, then the problem would not be as significant. However, reidentification can result in the identification of a large number of subjects. In our experiment we were able to reidentify 8,200 individuals out of the 11,000 subjects for which we had valid information. This amounts to about a 75% success rate. To achieve this high rate, we used a quasi-identifier made up of sex, inferred age, and death date for each victim. Using this quasi-identifier, we were able to match the victims from the Chicago Homicide Database to individuals contained in the Social Security Death Index.

Considering that the homicide database also had extra information regarding each case, such as flags marking certain records as having to deal with gang violence, homosexual lovers, or love triangles, it is evident that we identified individuals' private information. Although we were reidentifying victims, people who are now dead, the release of such information may still be damaging for the victims' families.

As we progressed through the course of our experiment, we were forced to ask ourselves if we were breaking any laws or regulations. An exhaustive search, unsurprisingly, found that there is little by way of any type of restriction on our actions. The medical field is currently trying to develop regulations specifying how databases containing patient data should be made anonymous, and only recently was a federal regulation put into place. However, as the legal analysis section noted, there is virtually no legislation nor judicial policy that covers other public data sets. In the absence of such policy, the public should voice concern and call for policy to be developed and implemented that: (a) regards reidentified data as equivalent to entrusted information, that is held to the same disclosure restrictions; (b) restricts companies from combining their databases containing entrusted information with publicly-available data sets; and, (c) applies deidentification standards, such as that just implemented on patient information data sets, to all other data sets.

Society must also recognize that legal policy alone is not enough. Technology to ensure anonymity of data must also be developed. Research, such as that being done by Latanya Sweeney, is beginning to address this need. Our own experiments illustrate that simple techniques beyond deidentification alone can decrease the ability to perform reidentification.

The time for action is now. Society must be made aware of the privacy concerns surrounding reidentification and how it can affect them. Currently, most of society is oblivious to the problem. As the amount of personal information continues to grow, society must ensure that the legal and technical restrictions are in place. Otherwise, any release of data will be subject to reidentification, and privacy will be a relic of the past.

References

[ACI] Appellate Court Of Illinois, "The Southern Illinoisan v. the Department of Public Health", No. 5-99-0568, 2001 Mar 28, [cited 2001 May 13], Available HTTP: http://www.state.il.us/court/Opinions/AppellateCourt/2001/5thDistrict/March/Html/5990568.htm

[BS] Buettner, Russ, and Sherman, William, "The 15 Most Sued Doctors in New York", New York Daily News, 2000 Mar 05, [cited 2001 May 10], Available HTTP: http://www.nydailynews.com/2000-03-05/News_and_Views/City_Beat/a-58900.asp

[CFR1] Code of Federal Regulations, "Confidentiality of Identifiable Research and Statistical Information", [cited 2001 May 15], Available HTTP: http://www.access.gpo.gov/nara/cfr/waisidx_00/28cfr22_00.html

[CFR2] Code of Federal Regulations, "Bureau of Prisons, Access to Records", [cited 2001 May 15], Available HTTP: http://www.access.gpo.gov/nara/cfr/waisidx_00/28cfr513_00.html

[CFR3] Code of Federal Regulations, "Department of Health and Human Services", [cited 2001 May 15], Available HTTP: http://www.access.gpo.gov/nara/cfr/waisidx_00/45cfrv1_00.html

[FSOG] Federal Statistical Office Germany, "Law on Statistics for Federal Purposes", 1987 Jan 22, [cited 2001 May 13], Available HTTP: http://www.statistik-bund.de/allg/e/ueber/bstatgueb_e.htm

[IORP] The Internet Open Records Project, "Dallas County Voting Records", [cited 2001 May 13], Available HTTP: http://www.openrecords.org/records/voting/dallas_voting/

[OLRC] U.S. House of Representatives, "Office of the Law Revision Counsel", [cited 2001 May 13], Available HTTP: http://law2.house.gov/

[SSO] Social Security Online, "Frequently Asked Questions", [cited 2001 May 10], Available HTTP: http://ssa-custhelp.ssa.gov

[SWE1] Sweeney, Latanya, "Computational Disclosure Control: A Primer on Data Privacy Protection", [cited 2001 May 13], Available HTTP: http://www.swiss.ai.mit.edu/classes/6.805/articles/privacy/sweeney-thesis-draft.pdf

[SWE2] Sweeney, Latanya, "Lecture 2: Data Explosion", [cited 2001 May 13], Available HTTP: http://sos.heinz.cmu.edu/dataprivacy/courses/dp1/lectures/lecture2paper.pdf

[SWE3] Sweeney, Latanya, "Lecture 3: Simple Demographics Identify People Uniquely", [cited 2001 May 13], Available HTTP: http://sos.heinz.cmu.edu/dataprivacy/courses/dp1/lectures/lecture3paper.pdf

[TDH] Texas Department of Health, "Information on Requests for Indexes", [cited 2001 May 13], Available HTTP: http://www.tdh.state.tx.us/bvs/registra/INDEX.HTM

Acknowledgements

Appendix A: Obtaining the SSDI Records

The SSDI interface available at RootsWeb.com restricts the user to viewing 15 records at a time. This limitation made it difficult to judge the completeness and uniqueness of the SSDI records.

In order to accurately gauge the appropriateness of the SSDI records for reidentification, we needed access to all of the records available for Chicago, Illinois from 1982 to 1995.

To this end, we implemented a simple web page spider in Perl using the LWP::UserAgent class. The program followed the flow chart at right. After each attempted page grab, the spider would pause for 15 seconds to avoid creating an unnecessary strain on the RootsWeb servers.

There were approximately 20,000 Chicago records per year, so the spider took about six hours to completely download a year of records. For 1982 to 1995, or 14 years, the total running time of the spider was about 3.5 days. There were 18.3 MB of downloaded records. The spider output these records to a tab-delimited file to simplify importing into a RDBMS.
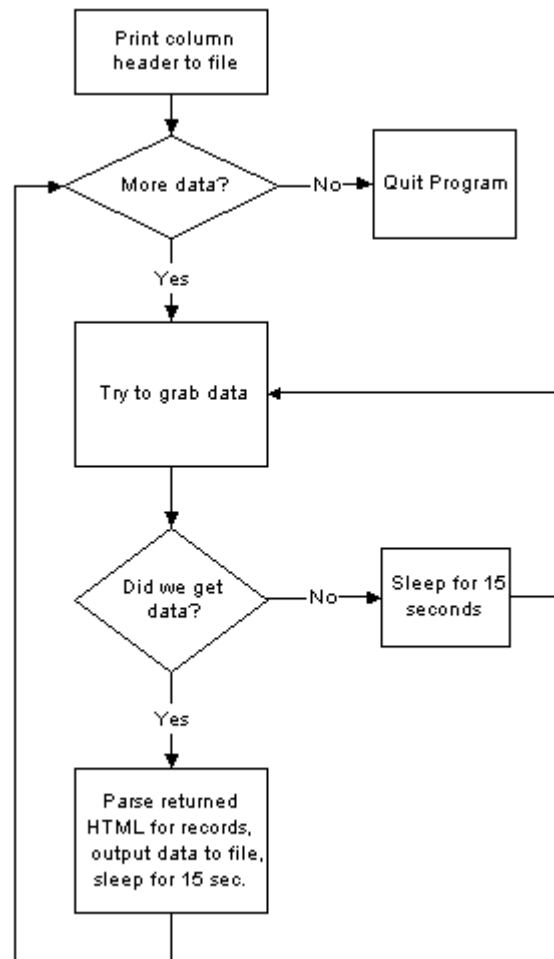

Figure 33: A Flow Chart of the SSDI Spider

Appendix B: SQL Queries

Query 1:

```sql
select last, middle, first, death_year,
       death_month, death_day, birth_year
from victim v, chicago_zip_codes, zip, ssdi_people s
where deathyr != 98 and place=1 and
      victim.centract = chicago_zip_codes.tract
      and zip.tract = chicago_zip_codes.tract
      and zip.num < 2 and vicage < 200
      and s.death_year = (1900 + v.deathyr)
      and s.death_month = v.deathmon
      and s.birth_year = (1900 + deathyr - vicage)
order by deathyr desc;
```

Query 2:

```sql
select last, middle, first, count(*), min(victim_id)
from victim v, ssdi_people s
where vicage < 200
      and s.death_year = (1900 + v.deathyr)
      and s.death_month = v.deathmon
      and s.birth_year = (1900 + deathyr - vicage)
group by last, middle, first
having count(*) = 1;
```

Query 3:

```sql
select last, middle, first, count(*), min(victim_id)
from victim v, ssdi_people s
where vicage < 200
        and s.death_year = (1900 + v.deathyr)
        and s.death_month = v.deathmon
        and ( ((s.birth_year = (1900 + deathyr - vicage))
              and ((death_month > birth_month)
                    or ((death_month = birth_month)
                        and (death_day >= birth_day))))
            or ((s.birth_year = (1900 + deathyr - vicage - 1))
              and ((death_month < birth_month)
                    or ((death_month = birth_month)
                        and (death_day <= birth_day)))))
group by last, middle, first
having count(*) = 1;
```

Query 4:

```sql
select s.last, s.middle, s.first, count(*), min(victim_id)
from victim v, ssdi_people s, first_names f
where vicage < 200 and
        f.first = s.first and f.gender = v.vicsex
        and s.death_year = (1900 + v.deathyr)
        and s.death_month = v.deathmon
        and ( ((s.birth_year = (1900 + deathyr - vicage))
              and ((death_month > birth_month)
                    or ((death_month = birth_month)
                        and (death_day >= birth_day))))
            or ((s.birth_year = (1900 + deathyr - vicage - 1))
              and ((death_month < birth_month)
                    or ((death_month = birth_month)
                        and (death_day <= birth_day)))))
group by s.last, s.middle, s.first
having count(*) = 1;
```