Foundations of Privacy Protection from a Computer Science Perspective

Latanya Sweeney
Center for Automated Learning and Discovery
School of Computer Science
H. John Heinz III School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA 15213 USA
latanya@andrew.cmu.edu

Abstract

In this paper, I explore the computational foundations for producing sufficiently anonymous data and define an *anonymous database system*, as one that makes individual and entity-specific data available such that the ability to identify individuals and other entities contained in the released data is controlled. While techniques for limiting and discovering inferences that can be drawn from released data can be adapted from work in statistical databases and from work in multi-level databases, anonymous databases differ in many significant ways. Here are a few differences: (1) all if not most of the data are released rather than a small sample; (2) the integrity of entity-specific details must be maintained rather than an overall aggregate statistic; and, (3) suppressing explicit identifiers, such as name and address, is not sufficient because combinations of other values, such as ZIP and birth date, can combine uniquely to re-identify entities.

I also provide a framework for reasoning about disclosure control and the ability to infer the identities of entities contained within data. I formally define and present *null-map*, *k-map* and *wrong-map* as models of protection. Each model provides protection by ensuring that released information maps to no, *k* or incorrect entities, respectively. I discuss the strengths and weaknesses of these protection models and provide real-world examples.

Contents at a Glance

1 Introduction

Society is experiencing exponential growth in the number and variety of data collections as computer technology, network connectivity and disk storage space becomes increasingly affordable. Data holders, operating autonomously and with limited knowledge, are left with the difficulty of releasing information that does not compromise privacy, confidentiality or national interests. In many cases the survival of the database itself depends on the data holder's ability to produce anonymous data because not releasing such information at all may diminish the need for the data, while on the other hand, failing to provide proper protection within a release may create circumstances that harm the public or others. Ironically, the broad availability of public and semi-public information makes it increasingly difficult to provide data that are effectively anonymous.

In this paper, I examine why it is so difficult to produce anonymous data in today's society and pose a framework for reasoning about solutions. The paper itself is divided into three main sections. I begin by looking at the nature of disclosure control problems and the identifiability of data. From there I compare and contrast a variety of protection techniques that are available. The paper ends with a formal presentation and examination of some protection models that attempt to effect disclosure control. Throughout the paper, I will present issues in the context of current disclosure control policies and systems.

Let me begin by being precise in my terminology and explain my use of medical privacy as a constant example. In general, I will discuss collections of information whose granularity of details are specific to an individual, a business, an organization or other entities and I term such collections, *entity-specific data*. If the entities represented in the data are individuals, then I may refer to the collection as *person-specific data*; however, even in these cases, the concepts being presented typically apply to broader collections of entity-specific data as well. By primarily using person-specific data and focusing on issues surrounding medical privacy, the motivations and risks often become transparent even though the underlying issues apply to many other kinds of data such as financial, statistical and national security information as well.

1.1 Tensions in releasing data

In the next two subsections, I look at different ways in which society has made decisions about sharing data, and I provide a way to reason about these findings. In the end, this examination motivates my use of

medical data as an example throughout this paper, even though the issues presented are not limited to medical data.

Quality versus anonymity

There is a natural tension between the quality of data and the techniques that provide anonymity protection. Consider a continuum that characterizes possible data releases. At one end of the continuum are person-specific data that are fully identified. At the other end are anonymous data that are derived from the original person-specific data, but in which no person can be identified. Between these two endpoints is a finite partial ordering of data releases, where each release is derived from the original data but for which privacy protection is less than fully anonymous. See Figure 1.

The first realization is that any attempt to provide some anonymity protection, no matter how minimal, involves modifying the data and thereby distorting its contents. So, as shown in Figure 1, movement along the continuum from the fully identified data towards the anonymous data adds more privacy protection, but renders the resulting data less useful. That is, there exists some tasks for which the original data could be used, but those tasks are not possible with the released data because the data have been distorted.

So, the original fully identified data and the derived anonymous data are diametrically opposed. The entire continuum describes the domain of possible releases. Framed in this way, a goal of this work is to produce an optimal release of data so that for a given task, the data remain practically useful yet rendered minimally invasive to privacy.

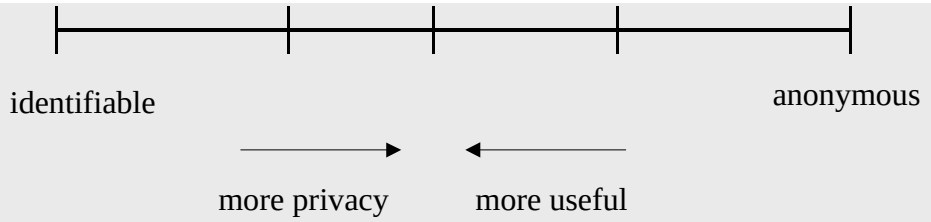


Figure 1 Optimal releases of data

Tug-of-war between data holders and recipients

The second realization that emerges from Figure 1 is that the usefulness of data is determined by the task to which the recipient puts the data. That is, given a particular task, there exists a point on the

continuum in Figure 1 that is as close to anonymous as possible, yet the data remain useful for the task. A release of data associated with that point on the continuum is considered optimal. In the next paragraphs, I provide a skeletal depiction of current practices that determine who gets access to what data. I show that the result can be characterized as a tug-of-war between data holders and data recipients.

In general, the practices of data holders and related policies do not examine tasks in a vacuum. Instead, the combination of task and recipient together are weighed against privacy concerns. This can be modeled as a tug-of-war between the data holder and societal expectations for privacy on one side, and the recipient and the recipient's use for the data on the other. In some cases such as public health legislation, the recipient's need for the data may overshadow privacy protections, allowing the recipient (a public health agent) to get the original, fully identified health data. See Figure 2 in which a tug-of-war is modeled. The privacy constraints on the data holder versus the recipient's demand for the data are graphically depicted by the sizes of the images shown. In the case illustrated, the recipient receives the original, fully identified data.

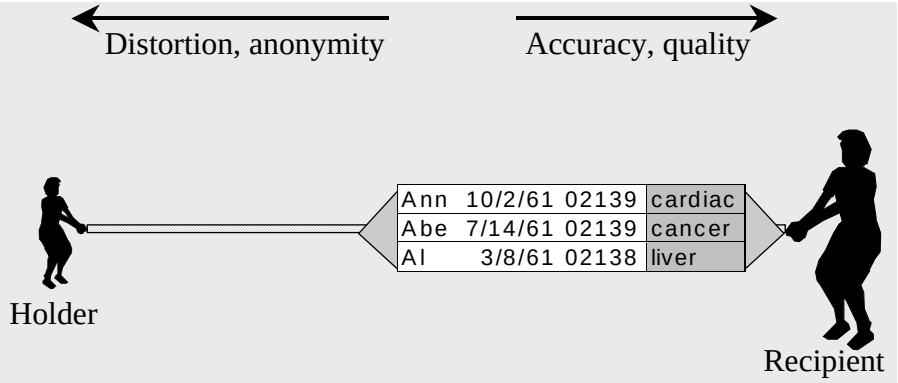


Figure 2. Recipient's needs overpower privacy concerns

Figure 3 demonstrates the opposite extreme outcome to that of Figure 2. In Figure 3, the data holder and the need to protect the confidentiality or privacy of the information overshadows the recipient and the recipient's use for the data and so the data is completely suppressed and not released at all. Data collected and associated with national security concerns provides an example. The recipient may be a news-reporting agent. Over time the data may eventually be declassified and a release that is deemed sufficiently anonymous provided to the press, but the original result is as shown in Figure 3, in which no data is released at all.

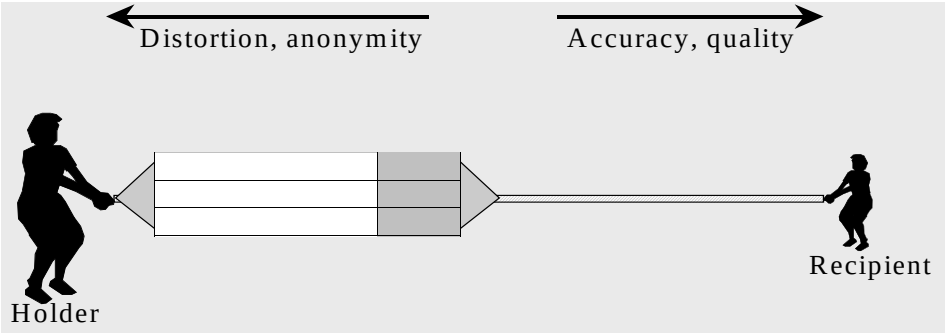


Figure 3 Data holder and privacy concerns overpower outside uses of the data

Figure 2 and Figure 3 depict situations in which society has made explicit decisions based on the needs of society as a whole. But secondary uses of medical data, for example, by marketing firms, pharmaceutical companies, epidemiological researchers and others do not in general lend themselves to such an explicit itemization. Figure 4 demonstrates situations in which the needs for privacy are weighed equally against the demand for the data itself. In such situations, a balance should be found in which the data are rendered sufficiently anonymous yet remain practically useful. As an example, this situation often occurs with requests by researchers for patient-specific medical records in which researchers seek to undertake clinical outcomes, or administrative research that could possibly provide benefits to society. At present, decisions are primarily based on the recipient receiving the original patient data or no data at all. Attempts to provide something in-between typically results in data with poor anonymity protection or data that is overly distorted. This work seeks to find ways for the recipient to get data that has adequate privacy protection, therefore striking an optimal balance between privacy protection and the data's fitness for a particular task.

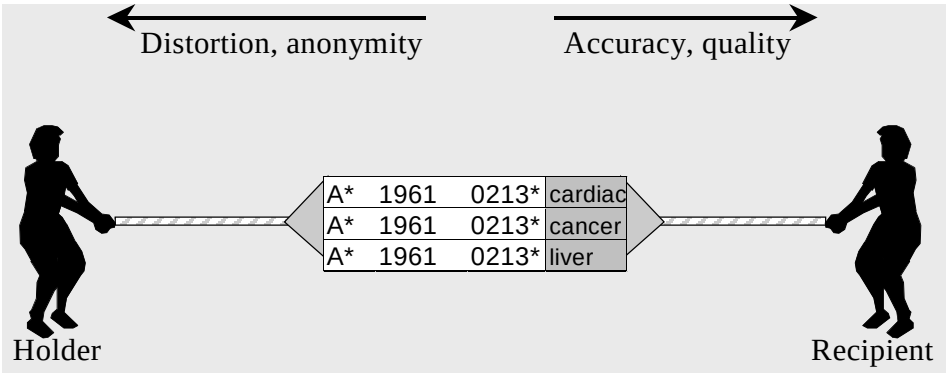


Figure 4. An optimal balance is needed between privacy concerns and uses of the data

At present, many data holders often make decisions arbitrarily or by ad hoc means. Figure 5 portrays the situation some state and federal agencies find themselves when they seek to produce public-use files for general use. Over the past few years, there has been a tremendous effort to make more data that is collected by government agencies available over the World Wide Web. In these situations, protecting the reputation of the agency, and the guarantees for privacy protection for which some agencies are legally bound, outweighs the demands of the recipient. In many of these cases, a strongly distorted version of the data is often released; the released data are typically produced with little or no consideration to the tasks required. Conversely, many other state and federal agencies release poorly protected data. In these cases, the individuals contained in the data can be easily re-identified. Examples of both of these kinds of released data are found in publicly and semi-publicly available hospital discharge data.

Neither way of releasing data yields optimal results. When strongly distorted data are released, many researchers cannot use the data, or have to seek special permission to get far more sensitive data than what are needed. This unnecessarily increases the volume of sensitive data available outside the agency. On the other hand, data that do not provide adequate anonymity may harm individuals.

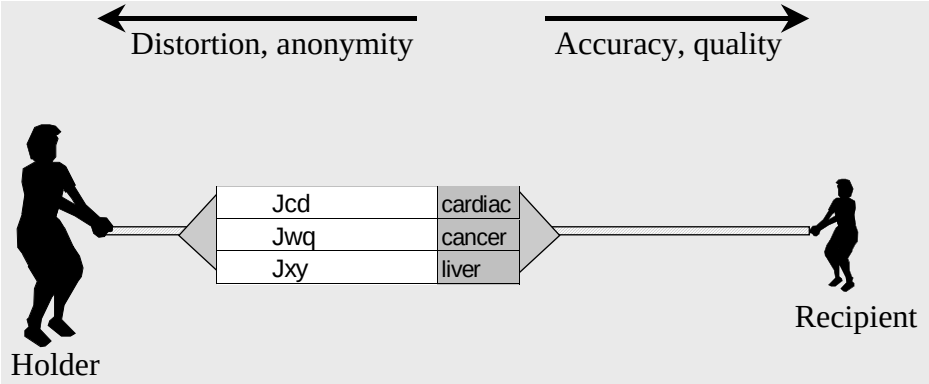


Figure 5. Data holder and privacy concerns limit uses of the data

In examining the different struggles between privacy and the sharing of person-specific data, I make the following claims:

Informal claim . Many current policies and practices support crude decisions. A recipient today too often receives the sensitive data itself, no data at all, overly distorted data that is of little or no use, or poorly protected data in which individuals can be re-identified.

Informal claim . Ultimately, the data holder must be held responsible for enforcing privacy protection because the data holder typically reaps a benefit and controls both data collection and dissemination.

While the claims above are independent of the content of data, the study of secondary uses of medical data in particular provides a natural incentive to find optimal solutions between researchers and data holders. After all, there are no legislative guidelines to empower one party so that it can overwhelm the other as was shown in Figure 2 and Figure 3. Also, state and federal agencies tend to be small in number and highly visible in comparison to the dramatic number of holders of medical data. Because there are so many holders of health data, it is hard to scrutinize their actions, and the resulting damage to individuals can be devastating yet hard to prove. And there exists strong financial incentives not to provide adequate protection in health data. On the other hand, research from data may lower health costs or save lives. For these reasons, focusing on the collection and sharing of medical data throughout this paper provides motivation for finding optimal releases of data and for integrating technology with policy for maximal benefit. Even though I focus on anonymity protection in medical data, the issues presented are just as pertinent to the confidentiality of businesses, governments and other entities in financial, marketing and other forms of data.

1.2 Introduction to privacy in medical data

I begin with some informal definitions. Identifiable personal health information refers to any information concerning a person's health or treatment that enables someone to identify that person. The expressions *personal health information* and *patient-specific health data* refer to health information that may or may not identify individuals. As I will show, in many releases of personal health information, individuals can be recognized. *Anonymous personal health information*, by contrast, contains details about a person's medical condition or treatment but the identity of the person cannot be determined.

In general usage, confidentiality of personal information protects the interests of the organization while privacy protects the autonomy of the individual; but, in medical usage, both terms often mean privacy.

Privacy protection and the Hippocratic oath

The historical origin and ethical basis of medical confidentiality begins with the Hippocratic Oath, which was written between the sixth century BC and the first century AD:

"Whatsoever I shall see or hear in the course of my dealings with men, if it be what should not be published abroad, I will never divulge, holding such things to be holy secrets."

Various professional associations worldwide reiterate this oath, and by pledging this oath, clinicians – licensed professionals such as doctors, nurses, pharmacists, radiologists, and dentists who access in the line of duty identifiable personal health information – assume the responsibility of securing this information. The resulting trust is the cornerstone of the doctor-patient relationship, allowing patients to communicate with their physicians and to share information regarding their health status. However, the doctor-patient *privilege* offers very little protection to patients regarding the confidentiality of their health information, being narrowly applicable to legal protection in some cases when a physician is testifying in court or in related proceedings.

Role of information technology

The role of information technology is critical to confidentiality. On the one hand, information technology offers comprehensive, portable electronic records that can be easily accessed on behalf of a given patient no matter where or when a patient may need medical care [1]. That very portability, on the other hand, makes it much easier to transmit quickly and cheaply records containing identifiable personal health information widely and in bulk, for a variety of uses within and among health care institutions and other organizations and agencies. The Office of Technology Assessment (OTA) found that current laws generally do not provide consistent or comprehensive protection of personal health information [2]. Focusing in the impact of computer technology, OTA concluded that computerization reduces some concerns about privacy of personal health information while increasing others.

Past policy efforts and computational disclosure control

Previous policy efforts to protect the privacy of personal health information were limited to decisions about who gets access to which fields of information. I examine in this paper four new computer systems that attempt to disclose information in such a way that individuals contained in the released data cannot be identified. These systems provide a spectrum of policy options. Decisions are no longer limited to who gets what information, but to how much generality or possible anonymity will exist in the released information.

Public concern over privacy

The public's concern about the confidentiality of personal health information is reflected in a 1993 poll conducted by Harris and Associates for Equifax. The results of the survey found that 96 percent of the respondents believe federal legislation should designate all personal health information as sensitive, and should impose severe penalties for unauthorized disclosure. Eighty percent of respondents were worried about medical record privacy, and 25 percent had personal experience of abuse related to personal health information [3].

A 1994 Harris-Equifax consumer privacy survey focused on how the American public feels about having their medical records used for medical research and how safeguards would affect their opinions about such systems and uses. Among a list of thirteen groups and organizations, doctors and nurses ranked first in terms of the percentage of Americans who were "very" confident (43 percent) that this group properly handled personal and confidential information. After hearing a description about how medical records are used by researchers to study the causes of disease, 41 percent of Americans surveyed said they would find it at least somewhat acceptable if their records were used for such research. Twenty-eight percent of those who initially opposed having their records used would change their position if a federal law made it illegal for any medical researcher to disclose the identity or any identifiable details of a person whose health records had been used. This would increase acceptance of this practice to over half those surveyed (58 percent) [4]. By extension, this survey implies strong public support for releases of personal health information in which persons contained in the information were unidentifiable.

Sharing medical data offers benefits to society

Analysis of the detailed information contained within electronic medical records promises many social advantages, including improvements in medical care, reduced institutional costs, the development

of predictive and diagnostic support systems [5], and the integration of applicable data from multiple sources into a unified display for clinicians [1]. These benefits, however, require sharing the contents of medical records with secondary viewers such as researchers, economists, statisticians, administrators, consultants, and computer scientists, to name a few. The public would probably agree that these secondary parties should know some of the information buried in the record, but such disclosure should not risk identifying patients.

Lots of medical data available from many sources

Beverly Woodward makes a compelling argument that, to the public, patient confidentiality implies that only people directly involved in one's health care will have access to one's medical records, and that these health professionals will be bound by strict ethical and legal standards that prohibit further disclosure [6]. The public is not likely to accept the notion that records are "confidential" if large numbers of people have access to their contents.

In 1996, the National Association of Health Data Organizations (NAHDO) reported that 37 states had legislative mandates to electronically gather copies of personal health information from hospitals [7] for cost-analysis purposes. Community pharmacy chains, such as Revco, maintain electronic records for over 60 percent of the 2.4 billion outpatient prescriptions dispensed annually. Insurance claims typically include diagnosis, procedure and medication codes along with the name, address, birth date, and SSN of each patient. Pharmaceutical companies run longitudinal studies on identified patients and providers. As more health maintenance organizations and hospitals merge, the number of people with authorized access to identifiable personal health information will increase dramatically because, as the National Research Council (NRC) recently warned, many of these systems allow full access to all records by any authorized person [8]. For example, assume a billing clerk at hospital X can view all information in all medical records within the institution. When hospital X merges with hospitals Y and Z, that same clerk may then be able to view all records at all three hospitals, even though the clerk may not need to know information about the patients at the other institutions.

Problems have been found

The NRC report also warns against inconsistent practices concerning releases of personal health information. If I approach a hospital as a researcher, I must petition the hospital's institutional review board (IRB) and state my intentions and methodologies; then the IRB decides whether I get data and in what form. But, if I approach the same hospital as an administrative consultant, data are given to me without IRB review. The decision is made and acted on locally.

Recent presentations by the secretary of the Department of Health and Human Services emphasize the threats to privacy stemming from misuse of personal health information [9]. There have been abuses; here are just a few:

- A survey of 87 Fortune 500 companies with a total of 3.2 million employees found that 35 percent of respondents used medical records to make decisions about employees [10].
- Cases have been reported of snooping in large hospital computer networks by hospital employees [11], even though the use of a simple audit trail – a list of each person who looked up a patient's record – could curtail such behavior [8].
- *Consumer Reports* found that 40 percent of insurers disclose personal health information to lenders, employees, or marketers without customer permission [12].

Abuses like the preceding underscore the need to develop safeguards.

This paper focuses on health data because: (1) the need to optimally produce health data releases that adequately protect privacy while still remaining practically useful is inherent in societal expectations, regulations and policies; (2) we are amidst an explosion in health data collection and sharing and these collections are not centralized and therefore act autonomously; and, (3) health data consists primarily of categorical values, which is characteristic of most new data collections.

1.3 All the data on all the people

Before I look at inference problems inherent in producing anonymous information, I first want to consider why concern over the problem appears to be escalating. There is currently unprecedented growth in the number and variety of person-specific data collections and in the sharing of this information. The impetus for this explosion has been the proliferation of inexpensive fast computers with large storage capacities operating in ubiquitous network environments.

In an attempt to characterize the growth in person-specific data, I introduced a metric named global disk storage per person or GDSP, which is measured in megabytes per person. GDSP is based on the total rigid disk drive space in megabytes of new units sold in a year divided by the world population in that year. Figure 6 uses GDSP figures to compute the amount of a person's time that can be documented on a page of text using a regularly spaced fixed font.

	1983	1996	2000
Storage space (TB)	90	160,623	2,829,288
Population (million)	4,500	5,767	6,000
GDSP (MB/person)	0.02	28	472
Time per page	2 months	1 hour	3.5 minutes

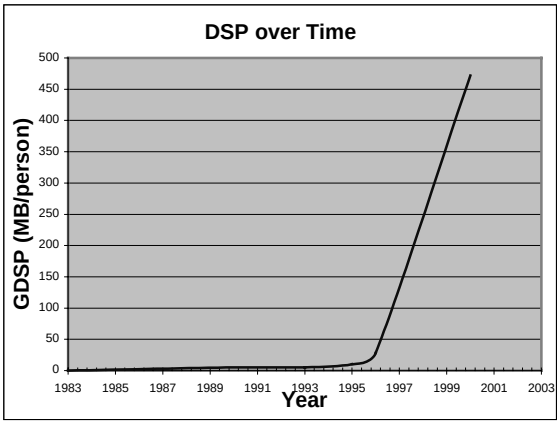


Figure 6 Global disk storage per person

In 1983 a half a page could be used to document each month of a person's life. These recordings included itemized long distance phone calls, credit card purchases, volume of electricity used, and so forth. In 1996, a page could be used to document each hour of a person's life. Recordings expanded in both size and number. Examples of new collections included items purchased at the grocery store, web sites visited, and the date and time in some locations a car proceeded through a tollbooth. In the year 2000, with 20 gigabyte drives leading the industry, it is projected that a page can be used to document every 3.5 minutes of a person's life. Collections are expanding to include biometric information such as, heart rate, pulse and temperature, video surveillance images and genetic information. One of the leading proponents of the information explosion is the health care industry, acting in the belief that having such information will help reduce cost and improve care.

Examples	1983	1996
Each birth	280	1864
Each hospital visit	0	663
Each grocery visit	32	1272

Figure 7 Estimated growth in data collections (per encounter) in Illinois (in bytes)

Figure 7 demonstrates how some data collections expanded from 1983 to 1996 for some person-specific encounters in the State of Illinois. The values are the number of *bytes* (letters, digits and other printable characters) that were stored for each person per encounter in the collection shown.

These examples exemplify recent behavioral tendencies recently found in the collection practices of person-specific data. These informally observed "trends" are enumerated below.

Behavior 1. Given an existing person-specific data collection, expand the number of fields being collected. I casually refer to this as the "*collect more*" trend.

Behavior 2. Replace an existing aggregate data collection with a person-specific one. I casually refer to this as the "*collect specifically*" trend.

Behavior 3. Given a question or problem to solve or merely provided the opportunity, gather information by starting a new person-specific data collection related to the question, problem or opportunity. I causally refer to this as the "*collect it if you can*" trend.

No matter how you look at it, all three tendencies result in more and more information being collected on individuals. Not only has there been a dramatic increase in the collection of person-specific data, but also in the sharing of collected data. I define four classes of access restrictions to person-specific data based on current practices. These are described in Figure 8.

Insiders only (Pr) *"private".*
Data collections that are available to authorized "insiders only" are considered to be privately held information because the only people who gain access are almost exclusively those who directly collected the information.

Limited Access (SPr) *"semi-private".*
Data collections denoted as having "limited access" are those where access extends beyond those who originally collected the information, but only an identifiable small number of people are eligible for access in comparison to a substantially larger number of people who are not eligible for access. This access policy typically includes an extensive application and review process.

Deniable Access (SPu) *"semi-public".*
Data collections having "deniable access" are those where an application and review process may exist but only an identifiable small number of people are denied access in comparison to a substantially larger number of people who are eligible for access.

No restrictions (Pu) *"public".*
Data collections having "no restrictions" are those where an application process may or may not exist, but the data collections are generally made available to all who request them.

Figure 8 Levels of access restrictions by data holders to person-specific data

There is no doubt that society is moving towards an environment in which society could have almost all the data on all the people. As a result, data holders are increasingly finding it difficult to produce anonymous and declassified information in today's globally networked society. Most data holders do not even realize the jeopardy at which they place financial, medical, or national security information when they erroneously rely on security practices of the past. Technology has eroded previous protections leaving the information vulnerable. In the past, a person seeking to reconstruct private information was often limited to visiting disparate file rooms and engaging in labor-intensive review of printed material in geographically distributed locations. Today, one can access voluminous worldwide public information using a standard handheld computer and ubiquitous network resources. Thus from seemingly innocuous anonymous data, and available public and semi-public information, one can often draw damaging inferences about sensitive information. However, one cannot seriously propose that all information with any links to sensitive information be suppressed. Society has developed an insatiable appetite for all kinds of detailed information for many worthy purposes, and modern systems tend to distribute information widely.

Primarily society is unaware of the loss of privacy and its resulting ramifications that stem from having so much person-specific information available. When this information is linked together it can provide an image of a person that can be as identifying as a fingerprint even if all explicit identifiers like

name, address, and phone number are removed. Clearly a loss of dignity, financial income and credit worthiness can result when medical information is widely and publicly distributed. A goal of the work presented in this paper is to control the release of data such that inferences about the identities of people and organizations and other sensitive information contained in the released data cannot be reliably made. In this way, information that is practically useful can be shared freely with guarantees that it is sufficiently anonymous and declassified. I call this effort the study of *computational disclosure control*.

In the next section, I introduce the basic problems of producing anonymous data.

2 Problems producing anonymous data

I now present examples that demonstrate why the problem of producing anonymous data is so difficult. Consider the informal definition of anonymous data below. While it is easy to understand what anonymous data mean, I will show by examples that it is increasingly difficult to produce data that are anonymous.

Definition (informal). anonymous data

> The term anonymous data implies that the data cannot be manipulated or linked to identify an individual.

A common incorrect belief is that removing all explicit identifiers from the data will render it anonymous; see the informal definition of de-identified data below. Many policies, regulations and legislation in the United States equate de-identified data and anonymous data. Drawing from experiments I conducted, I show in the next examples that de-identifying data provides no guarantee that the result is anonymous.

Definition (informal). de-identified data

> De-identified data result when all explicit identifiers such as name, address, and phone number are removed, generalized, or replaced with a made up alternative.

ZIP	Birth	Gender	Ethnicity
33171	7/15/71	m	Caucasian
02657	2/18/73	f	Black
20612	3/12/75	m	Asian

Figure 9 Data that look anonymous

Consider Figure 9. If I tell you that these three records are part of a large and diverse database then at first you may feel these three records are anonymous. If I subsequently tell you that the ZIP (postal code) 33171 consists primarily of a retirement community, then there are very few people of such a young age living there. The ZIP code 02657 is the postal code for Provincetown, Massachusetts and reportedly there are only five black women who live there year round. Likewise, 20612 may have only one Asian family. In each of these cases it was information outside the data that helped re-identify individuals.

Diagnosis	Diagnosis date	ZIP
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…

Figure 10 Cancer registry that looks anonymous

Recently, a state Department of Public Health received a Freedom of Information request from a newspaper that was researching occurrences of a rare cancer in a small region of the state. Although the paper only wanted diagnosis, date of diagnosis (month, day and year) and ZIP code (5 digits) for each patient in question, the state refused claiming that sensitive information might be gleamed from these data. In an attempt to discover how anonymous such information in question could be, I conducted an experiment. Within a few hours the name, and in some cases the Social Security number of five out of five patients submitted were accurately identified using only publicly available information. Further, four of the five cases had a diagnosis of Kaposi's Sarcoma which when found in young men is an indicator of AIDS and revealing such may have been prohibited by state law. Figure 10 shows an example of this data schema. A more extensive re-identification experiment, using similar data and achieving similar results was performed. It is difficult to believe that such seemingly innocuous information can be so easily re-identified.

- Patient **ZIP Code**
- Patient **Birth Date**
- Patient **Gender**
- Patient Racial Background
- Patient Number
- Visit Date
- Principal Diagnosis Code (ICD9)
- Procedure Codes (up to 14)
- Physician ID#
- Physician ZIP code
- Total Charges

Figure 11 Attributes often collected statewide

I will now demonstrate how linking can be used to perform such re-identification. The National Association of Health Data Organizations (NAHDO) reported that 37 states have legislative mandates to collect hospital level data and that 17 states have started collecting ambulatory care data from hospitals, physicians offices, clinics, and so forth [13]. Figure 11 contains a subset of the fields of information, or attributes, that NAHDO recommends these states accumulate. The few attributes listed in Figure 12 include the patient's ZIP code, birth date, gender, and ethnicity. Clearly, the data are de-identified. The patient number in earlier versions was often the patient's Social Security number and in subsequent versions was a scrambled Social Security number [14]. By scrambled I mean that the digits that compose the Social Security number are moved around into different locations. If a patient's record is identified and their Social Security number known, then the scrambling algorithm can be determined and used to identify the proper Social Security numbers for the entire data set.

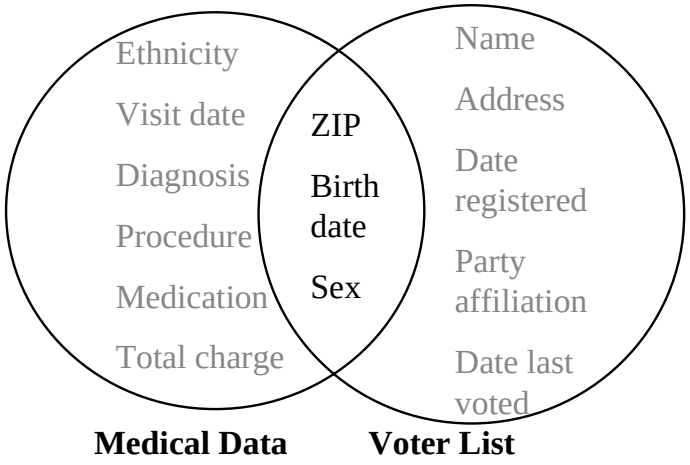


Figure 12 Linking to re-identify data

For twenty dollars I purchased the voter registration list for Cambridge Massachusetts and received the information on two diskettes [15]. Figure 12 shows that these data included the name, address, ZIP code, birth date, and gender of each voter. This information can be linked using ZIP code, birth date and gender to the medical information described in Figure 11, thereby linking diagnosis, procedures, and medications to particularly named individuals. The question that remains of course is how unique would such linking be.

As I reported previously, the 1997 voting list for Cambridge Massachusetts contained demographics on 54,805 voters. Of these, birth date, which is the month, day, and year of birth, alone could uniquely identify the name and address of 12% of the voters. One could identify 29% of the list by just birth date and gender; 69% with only a birth date and a five-digit zip code; and 97% when the full postal code and birth date were used. Notice that these are only one and two way combinations and do not include three way combinations or beyond. These values are summarized in Figure 13.

Attribute Combinations	Uniqueness
Birth date alone (mm/dd/yr)	12%
Birth date and gender	29%
Birth date and 5-digit ZIP	69%
Birth date and full postal code	97%

Figure 13 Value uniqueness in voter list

In general I can say that the greater the number and detail of attributes reported about an entity, the more likely that those attributes combine uniquely to identify the entity. For example, in the voter list, there were 2 possible values for gender and 5 possible five-digit ZIP codes; birth dates were within a range of 365 days for 100 years. This gives 365,000 unique values, but there were only 54,805 voters. My recent experiments reveal that about 87% of the population of the United States can be uniquely identified using {*date of birth*, *5-digit ZIP*, *gender*}.

Consider the following example. In Massachusetts the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. GIC collected de-identified medical encounter level 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 reported giving a copy of the data to researchers and selling a copy to industry [16]. William Weld was governor of Massachusetts at that time and his medical records were in that data. Governor Weld lived in Cambridge Massachusetts. Six people had his particular birth date. Only three of them were men and he was the only one in his five-digit zip code.

Clearly the risks of re-identifying data depend both on the content of released data and on other related information. Most municipalities and states sell population registers such as voter lists, local census data, birth records and motor vehicle information. There are other sources of population registers such as trade and professional association lists. Such information can often be uniquely linked to de-identified data to provide names, addresses, and other personal information.

These real-world examples demonstrate two major difficulties in providing anonymous data: (1) knowledge a viewer of the data may hold or bring to bear on the data is usually not known beforehand by the data holder at the time of release; and, (2) unique and unusual values and combinations of values appearing within the data themselves often makes identification of related entities easier. The examples also underscore the need to develop solutions that limit the ability to link external information to data and therefore control the inferences that can be drawn.

The outline for the remainder of this paper is as follows. In the next section, section 3, I discuss related work. I then survey disclosure control techniques and the nature of disclosure control in section 4. A formal presentation with accompanying definitions of protection models is presented in section 5. Finally, four systems are presented and compared in section 6.

3 Related Work

The problem of controlling inferences that can be drawn from released data is not new. There are existing works in the statistics community on statistical databases and in the computer security community on multi-level databases to consider. These are described here briefly. Before examining these traditions, I establish a common vocabulary by adopting the following definitions.

Unless otherwise stated, the term *data* refers to entity-specific information that is conceptually organized as a table of rows (or records) and columns (or fields). Each row is termed a *tuple*. A tuple contains a relationship among the records or set of values associated with an entity. Tuples within a table are not necessarily unique. Each column is called an *attribute* and denotes a field or semantic category of information that is a set of possible values; therefore, an attribute is also a domain. Attributes within a table are unique. So by observing a table, each row is an ordered n-tuple of values $<d_1, d_2, \ldots, d_n>$ such that each value d_j is in the domain of the j-th column, for j=1, 2, …, n where n is the number of columns. In mathematical set theory, a relation corresponds with this tabular presentation, the only difference is the absence of column names. Ullman provides a detailed discussion of relational database concepts [17].Throughout the remainder of this paper each tuple is assumed to be specific to one entity and no two tuples pertain to the same entity. This assumption simplifies discussion without loss of applicability.

To draw an *inference* is to come to believe a new fact on the basis of other information. A *disclosure* means that explicit or inferred information about an entity was released that was not intended. This definition may not be consistent with colloquial use but is used in this work consistent with its meaning in statistical disclosure control. So, *disclosure control* attempts to identify and limit disclosures in released data. Typically the goal of disclosure control with respect to person-specific data is to ensure that released data are anonymous.

3.1 Statistical databases

Federal and state statistics offices around the world have traditionally been entrusted with the release of statistical information about all aspects of the populace [18]. The techniques, practices and theories from this community however, have historically had three tremendous advantages. First, most statistics offices held centralized, sole-source exhaustive collections of information and therefore could often determine the sensitivity of many values using their data alone. Second, statistics offices primarily produced summary data, which by the nature of aggregation could often hide entity-specific information though care still had to be taken to protect against inferences. Third, many statistical offices worked with data that primarily consisted of attributes having continuous values, but he recent surge in data has been in attributes having categorical values. Fourth, some form of all if not most of the data are released rather than a sample. Finally, statistics offices previously released information in an environment whose computational power and access to other data was extremely limited. These advantages have been eroded in today's environment. Today's producers of useful publicly available data must contend with autonomous releases of entity-specific information by other data holders and with recipients who are technologically empowered.

3.2 Multi-level databases

Another related area is aggregation and inference in multi-level databases [19, 20, 21, 22, 23, 24] which concerns restricting the release of lower classified information such that higher classified information cannot be derived. Denning and Lunt [25] described a multilevel relational database system (MDB) as having data stored at different security classifications and users having different security clearances.

Many aggregation inference problems can be solved by database design [26, 27], but this solution is not practical in the entity-specific data setting described in sections 1 and 2. In today's environment, information is often divided and partially replicated among multiple data holders and the

data holders usually operate autonomously in making disclosure control decisions. The result is that disclosure control decisions are typically made locally with incomplete knowledge of how sensitive other holders of the information might consider replicated data. For example, when somewhat aged information on joint projects is declassified differently by the Department of Defense than by the Department of Energy, the overall declassification effort suffers; using the two partial releases, the original may be reconstructed in its entirety. In general, systems that attempt to produce anonymous data must operate without the degree of omniscience and level of control typically available in the traditional aggregation problem.

In both aggregation and MDB, the primary technique used to control the flow of sensitive information is *suppression*, where sensitive information and all information that allows the inference of sensitive information are simply not released [28]. Suppression can drastically reduce the quality of the data, and in the case of statistical use, overall statistics can be altered, rendering the data practically useless. When protecting national interests, not releasing the information at all may be possible, but the greatest demand for entity-specific data is in situations where the data holder must provide adequate protections while keeping the data useful, such as sharing person-specific medical data for research purposes. In section 4 and beyond, I will present other techniques and combinations of techniques that produce more useful data than using suppression alone.

3.3 Other areas

Denning [29] along with others [30] and Duncan and Mukherjee [31] were among the first to explore inferences realized from multiple queries to a database. For example, consider a table containing only (*physician*, *patient*, *medication*). A query listing the patients seen by each physician, i.e., a relation R(*physician*, *patient*), may not be sensitive. Likewise, a query itemizing medications prescribed by each physician may also not be sensitive. But the query associating patients with their prescribed medications may be sensitive because medications typically correlate with diagnoses. There exist many kinds of situations in which the sensitive, unreleased information, can be inferred from the other two. The common solution to this problem involves suppressing most or all of the data, even when inferences are restricted to the attributes and values contained within the data. In contrast the work presented in this paper poses real-time solutions to this problem by advocating that the data be first rendered sufficiently anonymous, and then the resulting data used as the basis on which queries are processed. In sections 4, 5 and 6, I show ways data can be made sufficiently anonymous such that the queries may still be useful.

In summary, the catalyst for now examining disclosure control in a broader context has been the dramatic increase in the availability of entity-specific data from autonomous data holders. Having so much data readily available has expanded the scope and nature of inference control problems and exasperated established operating practice. A goal of this paper is to shed light on these problems and to provide comprehensive models for understanding, evaluating and constructing computational systems that control inferences in this setting.

4 A framework for reasoning about disclosure control

In section 5 a formal presentation is provided and in section 6 real-world systems are evaluated, but first, in this section, I provide a framework for reasoning about disclosure control and I survey some disclosure limitation techniques using this framework.

4.1 Survey of disclosure limitation techniques

I begin by first introducing commonly employed disclosure limitation techniques; Figure 14 contains a listing. Here is a quick description of each technique though some were introduced earlier. De-identification was described in an earlier informal definition (on page 16). Suppression was introduced in section 3.2 (see page 22). *Encryption* is a process of making values secret by replacing one value with another in such a way that certain properties with respect to reversing the process are maintained. *Swapping values* involves exchanging the values associated with an attribute in two tuples where the value from the first tuple becomes the value for the second and vice versa. *Generalization* replaces a value with a more general, less specific alternative. *Substitution* replaces a value with another value in its equivalence class. *Sampling* restricts the number of tuples that will be released. *Scrambling* is a reordering of tuples and is used when the order of appearance of tuples in a release allows inference[1]. Changing *outliers to medians* requires detecting unusual values and replacing them with values that occur more commonly. *Perturbation* involves making changes to values, usually to maintain some overall aggregate statistic. *Rounding* is often used on continuous variables to group values into ranges. Adding *additional tuples* dilutes the number of tuples containing *real* information but values within the newly generated tuples can be chosen to maintain certain aggregate properties. *Additive noise* involves the random incrementing or decrementing of values.

[1] This is slightly inconsistent with the relational model, but in practical use is often an issue.

Value and Attribute Based	De-identification	Substitution
	Suppression	Outlier to medians
	Encryption	Perturbation
	Swap values	Rounding
	Generalize values	Additive noise
Tuple based	Sampling	
	Add tuples	
	Scramble tuples	
Other	Query restriction	
	Summaries	

Figure 14 Disclosure limitation techniques

Query restriction and *summary data* described earlier are not disclosure limitation techniques but rather special circumstances in which disclosure control is required. In summary data and query restriction, values are often suppressed so as not to reveal sensitive information. This work poses a solution to many problems in query restriction and summarizing by basing queries and summaries on data released from data already determined to be sufficiently anonymous or in the process of being rendered sufficiently anonymous during the query process itself.

Notice that all of these techniques have the advantage that a recipient of the data can be told what was done to the data in terms of protection. For data to be useful and results drawn from data to be properly interpreted, it is critical to share what techniques and associated parameters were employed in protecting the confidentiality of entities within the data. Of course usefulness is determined from the point of view of a recipient of the data and what is useful to one recipient is not necessarily beneficial to another. For example, perturbation can render data virtually useless for learning entity-specific information from the data or identifying entity-specific correlation. On the other hand, suppression can render data virtually useless for statistical purposes.

During the application of any technique, decisions must be made and these decisions can dramatically impact the data's fitness for a particular purpose. For example, consider a situation in which it is necessary to suppress either values associated with the attribute *ethnicity* or those associated with the attribute *ZIP*. If the recipient of the data is an epidemiologist studying cancer rates near toxic waste sites, then the suppression of *ZIP* may render the data useless. Conversely, if the epidemiologist was studying the prevalence of heart disease among various ethnic groups, then the suppression of *Ethnicity* may have the same ill result. Notice that the data holder cannot release both versions, because doing so may allow the two releases to be linked and reveal all information. Data holders must typically decide a priori for which uses released information will be best suited.

4.2 Reasoning about disclosure control

The goal of this section is to provide a framework for constructing and evaluating systems that release information such that the released information limits what can be revealed about properties of the entities that are to be protected. For convenience, I focus on person-specific data and the property to be protected is the identity of the subjects whose information is contained in the data. A disclosure implies that an identity was revealed. Consider the informal definition below. Basically, an anonymous data system seeks to effect disclosure control. I use the framework presented in this section to describe the requirements of an anonymous data system and in section 5 I formally define such.

Definition (informal). anonymous data system

> An anonymous data system is one that releases entity-specific data such that particular properties, such as identity, of the entities that are the subject of the data are not released.

I can be more specific about how properties are selected and controlled. Recall the real-world examples provided in section 2. In those cases, the need for protection centered on limiting the ability to link released information to other external collections. So the properties to be controlled are operationally realized as attributes in the privately held collection. The data holder is expected to identify all attributes in the private information that could be used for linking with external information. Such attributes not only include explicit identifiers such as name, address, and phone number, but also include attributes that in combination can uniquely identify individuals such as birth date and gender. The set of such attributes has been termed a *quasi-identifier* by Dalenius [32] and an *identificate* by Smith [33]. So operationally, an anonymous data system releases entity-specific data such that the ability to link to other information using the quasi-identifier is limited.

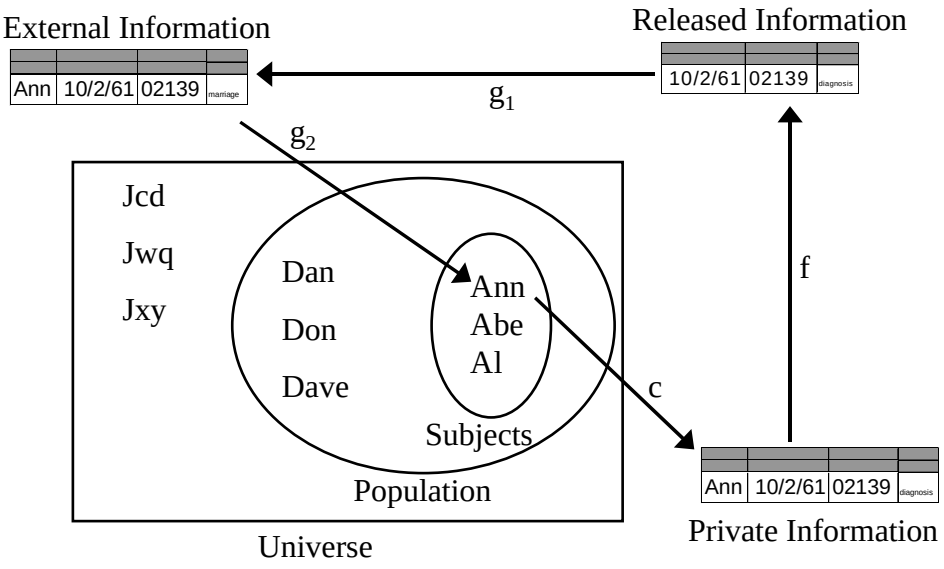


Figure 15 Release using de-identification

Figure 15 provides an overview of the disclosure control process. Population consists of persons who are identified as {*Dan*, *Don*, *Dave*, *Ann*, *Abe*, *Al*}, A subset of Population called Subjects is the set of people, in this case, {*Ann*, *Abe*, *Al*}, whose information appears in PrivateInformation. Universe consists of Population and the set of *pseudo-entities* {*Jcd*, *Jwq*, *Jxy*}. Pseudo entities are not considered real individuals, as are the members of Population. Instead, the existence of a pseudo-entity is implied by a set of values, which are associated with attributes that identify people, when in fact no such person is associated with that particular set of values.

There exists a collection function *c*: Subjects → PrivateInformation that maps information about members of Subjects into PrivateInformation. The function *f* is a disclosure limitation function such that *f*: PrivateInformation → ReleasedInformation. In the example shown in Figure 15, *f* simply de-identifies tuples from PrivateInformation; and so, the explicit identifier *Ann* is not found in ReleasedInformation.

ExternalInformation results from joining all publicly (and semi-publicly) available information. The relations g_1 and g_2 illustrate how a tuple in ReleasedInformation can be linked to a tuple in ExternalInformation to re-identify *Ann*, the original subject. *The problem of producing anonymous information can be described as constructing the function f such that some desired invariant exists or some specific assertion can be made about g_1 and g_2. Such an invariant or assertion forms the basis for protection.*

In the example shown in Figure 15, the function *f* is simply the de-identification function and the functions g_1 and g_2 show that *f* is not sufficient; it allows a disclosure. Therefore, merely suppressing explicit identifiers is inadequate.

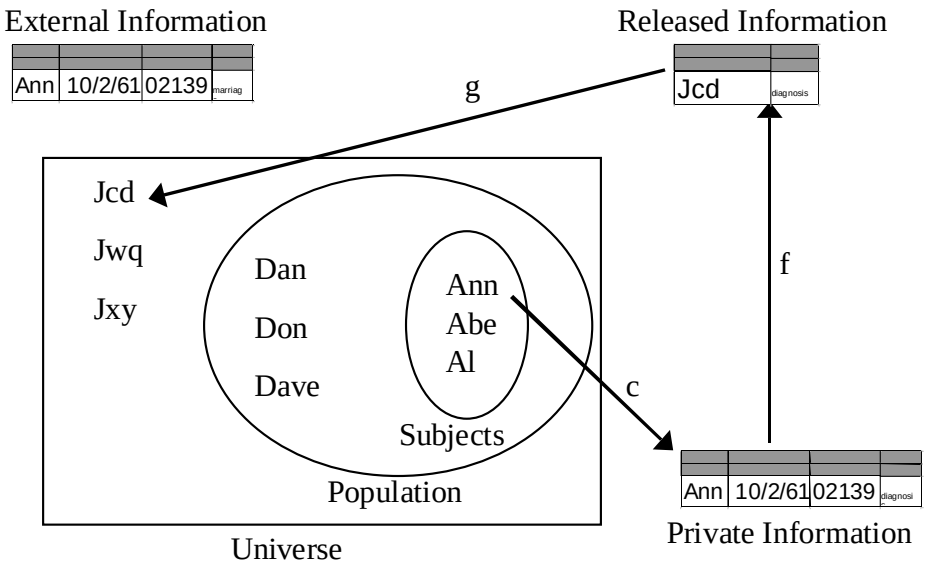


Figure 16 Release using encryption

Consider Figure 16. The function *f* seeks to protect the entire quasi-identifier {*name*, *birth date*, *ZIP*} by simply encrypting the associated values. If strong encryption is used and the encrypted values are not used with other releases, then as the diagram in Figure 16 illustrates, the relation *g* will map to a pseudo-entity, being unable to link to ExternalInformation. If on the other hand, *f* used weak encryption then the relation *g* would be able to map directly to *Ann* by simply inverting *f*. Using this approach with strong encryption clearly provides adequate protection, but such protection is at the cost of rendering the resulting information of limited use. Similar results are realized if *f* involved suppression rather than encryption. As shown in Figure 16, the only attribute that remains practically useful is *diagnosis* with no consideration to age or geographical location.

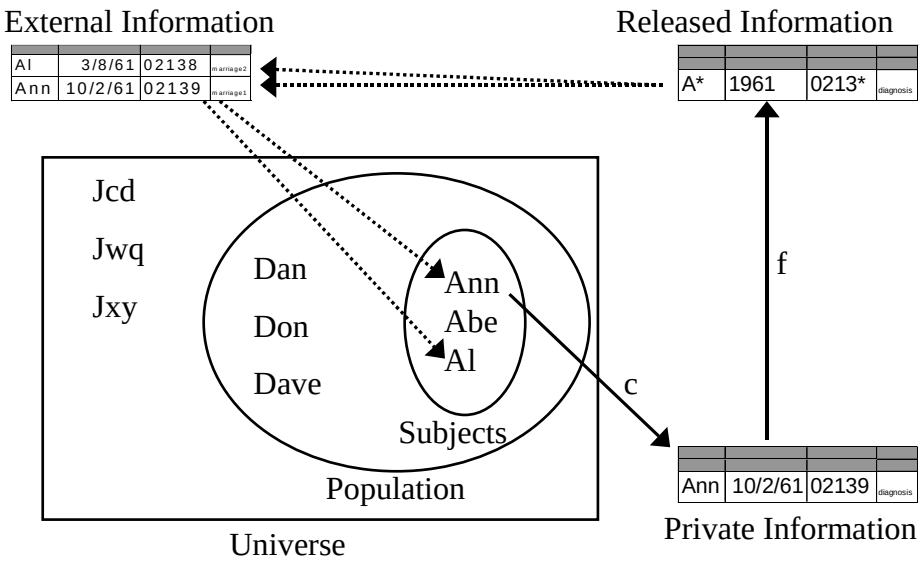


Figure 17 Release using generalization

Consider Figure 17. The function *f* generalizes the attributes of the quasi-identifier. I will take a moment to first talk about what is meant by generalizing an attribute and then I will return to this scenario for disclosure limitation.

The idea of generalizing an attribute is really a simple concept. A value is simply replaced by a less specific, more general value that is faithful to the original value. In Figure 18 the original ZIP codes {*02138*, *02139*} can be generalized to *0213**, thereby stripping the rightmost digit and semantically indicating a larger geographical area. Likewise {*02141*, *02142*} are generalized to *0214**, and {*0213**, *0214**} could be further generalized to *021***.

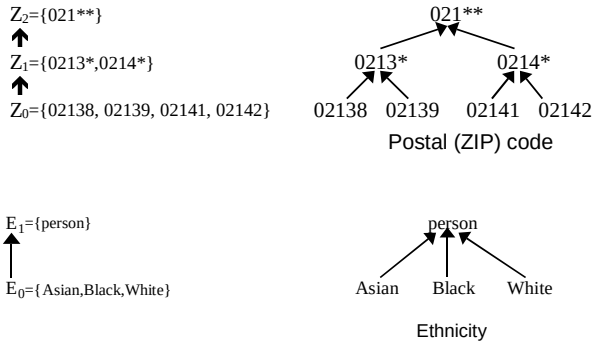


Figure 18 Generalizing an attribute

Generalization is effective because substituting values with their more generalized values increases the number of tuples having the same values. The single term requirement on the maximal element insures that all values associated with an attribute can eventually be generalized to a single value. In an earlier work, I demonstrated that all values of all attributes can be semantically organized into generalization hierarchies. Notice in Figure 18 that the values {*Asian*, *Black*, *White*} generalize to *Person*. This means that a generalization of an *Ethnicity* attribute given this hierarchy is similar to suppressing the entire attribute, thereby demonstrating that generalizing an attribute to its maximal element provides almost the same protection and distortion as suppressing the attribute. The relationship between generalization and suppression will be further discussed in section 6.4.

I now return to Figure 17. The disclosure limitation function *f* generalizes the attributes of the quasi-identifier to produce ReleasedInformation. Tuples in ReleasedInformation can then be linked to ExternalInformation ambiguously. In Figure 17, the tuple shown in ReleasedInformation links to both *Al* and *Ann* in ExternalInformation and so, it relates back to both of them in Subjects. The disclosed *diagnosis* cannot be confidently attributed to either *Al* or *Ann*. In fact, a *k* can be chosen such that *f* generalizes tuples from PrivateInformation in such a way that there are at least *k* possible entities to which each released tuple may refer. Additional protection can often be realized when tuples in ReleasedInformation are ambiguously linked to tuples in ExternalInformation such that the resulting identifications do not only refer to entities in Subjects but also refer to other entities in Universe that are not in Subjects.

A problem however is choosing the right size for *k*. It is based on several parameters including direct and economical communication connections to Subjects. Here is an example. I reviewed some archives from old email exchanges on a newsgroup list and found a couple of email messages pertaining to a chance encounter in Cambridge, Massachusetts between a young woman, who I will call Alice, and a young man, who I will call Bob. During the brief conversation between Alice and Bob, no names, addresses or phone numbers were exchanged. Several days later Alice engaged in an email exchange on a newsgroup list in which she provided a casual description of Bob. I constructed a composite of Bob from the email messages. Here is an overview of the details. Bob was about 5'8" in height with dark features. His parents were from Greece. He was believed to live near the water, to enjoy playing soccer and to be an MIT graduate student in electrical engineering or computer science. Given this basic description, I sent a single email message to all members of the electrical engineering and computer science department at MIT. Approximately 1,000 people could have received the message. Five replies were received. All of them had one name, which turned out to be the correct individual. The man himself was quite shocked because he had merely had a private conversation carried in a personal situation and he had not even given his name, phone number, or address. With respect to this disclosure control model, *k*

would be about 100 in this case and still that was not sufficient because of the direct and economical communication connection to all-possible subjects and sources of additional information.

This concludes my survey of disclosure limitation techniques and introduction of this framework for reasoning about disclosure control. In the next section I introduce formal models of protection. Following that, I compare and contrast some real-world systems.

5 Formal protection models

In this section, I formally bring the pieces together; namely, the lessons learned in the real-world examples of section 2, the issues presented in the discussion of related work in section 3 and the framework for reasoning about disclosure control that was presented in section 4. Terms mentioned casually and defined informally earlier will now be presented formally. So, I begin this section by formally defining the terms I have been using, leading up to the definition of a basic anonymous data system termed $\mathbf{ADS_0}$. From there, I introduce basic protection models termed *null-map*, *k-map* and *wrong-map* which provide protection by ensuring that released information maps to no, *k* or incorrect entities, respectively. The non-technical reader may elect to skip this section altogether and continue with section 6 (on page 40), which examines four real-world systems that attempt to effect disclosure control.

As stated earlier in section 3, I assume the classical relational model of databases. The definition below defines a table and attributes consistent with this model.

Definition. attributes

Let $\mathsf{B}(A_1,\ldots,A_n)$ be a *table* with a finite number of tuples. The finite set of *attributes* of B are $\{A_1,\ldots,A_n\}$.

Given a table $\mathsf{B}(A_1,\ldots,A_n)$, $\{A_i,\ldots,A_j\} \subseteq \{A_1,\ldots,A_n\}$, and a tuple $t \in \mathsf{B}$, I use $t[A_i,\ldots,A_j]$ to denote the sequence of the values, $v_i,\ldots,v_j$, of $A_i,\ldots,A_j$ in t. I use $\mathsf{B}[A_i,\ldots,A_j]$ to denote the projection, maintaining duplicate tuples, of attributes $A_i,\ldots A_j$ in B.

Definition. entity

Let p_i = { (A_i, v_i) : A_i is an attribute and v_i is its associated value}. I say p_i is an *entity*. U = {p_i : p_i is an entity} is a finite set I term a *population of entities*.

Definition. collection function

Given a population of entities U and a table T, I say f_c is a collection function on U. That is, f_c: $\mathsf{U} \rightarrow \mathsf{T}$ is a *collection function* and T is an *entity-specific table*. I say that T is a *person-specific table* if the entities are people.

If T is an entity specific table containing information about entities in U and T contains no additional tuples, then each tuple in T corresponds to information on at least one entity in U. This is memorialized in the following theorem.

Theorem.

Given a population of entities U, a table $\mathsf{T}(A_1,\ldots,A_n)$, a collection function f_c: $\mathsf{U} \rightarrow \mathsf{T}$, and $\{A_i,\ldots,A_j\} \subseteq \{A_1,\ldots,A_n\}$:

f_c is onto $\Rightarrow \forall t[A_i,\ldots,A_j] \in \mathsf{T}, \exists p_i \in \mathsf{U}$ such that $\forall (A_x, v_x) \in p_i$ where $A_x \in \{A_i,\ldots,A_j\}$ and $v_x = t[A_x]$.

Proof.

By definition, a function f_c from U to T is onto (or a surjection) if and only if for every element in $t \in \mathsf{T}$ there is an element $p \in \mathsf{U}$ with $f_c(p)=t$.

Example.

Let T be a table of visits to a hospital emergency room. Let U reflect the population of people within the geographical area serviced by the hospital. Then, f_c: $\mathsf{U} \rightarrow \mathsf{T}$ is the process for recording hospital visits. Notice that f_c is the collection function and f_c is onto.

Definition. disclosure control function

Given a table T and a finite set of tables B, I say f is a disclosure control function on {T}. That is, f: {T} $\rightarrow$ B is a *disclosure control function* .

Definition. re-identification relation

Given a population of entities U, an entity-specific table T and f_c: U $\rightarrow$ T,

I say f_g is a *re-identification relation* if and only if:

$\exists p_i \in$ U such that $p_i \in f_g(f_c(p_i))$ and $|f_g(f_c(p_i))| = k$, where $1 \leq k << |U|$.

I also say that f_g is a re-identification of p_i and I say that f_g uniquely identifies p_i if k=1.

Pseudo entities are not real entities but their existence is implied by a set of values, one or more of which are false, that are associated with attributes that seem to identify them as entities. This is described in the definition below.

Definition. pseudo-entities

Given a population of entities U, an entity-specific table T, f_c: U $\rightarrow$ T and a re-identification relation f_g: T $\rightarrow$ U' where U $\subseteq$ U'. I say (U'-U) is the finite set of *pseudo-entities*.

The following definition formally introduces a quasi-identifier, which, as was discussed earlier (on page 25), is a set of attributes whose associated values may be useful for linking to re-identify the entity that is the subject of the data.

Definition. quasi-identifier

Given a population of entities U, an entity-specific table T, f_c: U $\rightarrow$ T and f_g: T $\rightarrow$ U', where U $\subseteq$ U'. A quasi-identifier of T, written Q_T, is a set of attributes $\{A_i,\ldots,A_j\} \subseteq \{A_1,\ldots,A_n\}$ where:

$\exists p_i \in$ U such that $f_g(f_c(p_i)[Q_T]) = p_i$.

Example.

Let V be the voter-specific table described in Figure 12 as the voter list. A quasi-identifier for V, written Q_V, is {*name*, *address*, *ZIP*, *birth date*, *gender*}.

Linking the voter list to the medical data as shown in Figure 12, clearly demonstrates that {*birth date*, *ZIP*, *gender*} $\subseteq Q_V$. However, {*name*, *address*} $\subseteq Q_V$ because these attributes can also appear in external information and be used for linking.

The goal of disclosure control is to limit the extent to which released information can be confidently linked to other available information. In the case of anonymity, it is usually publicly available data on which linking is to be prohibited and so attributes which appear in privately held data and also appear in publicly available data are candidates for linking; therefore, these attributes constitute the quasi-identifier and the disclosure of these attributes must be controlled. It is believed that the data holder can easily identify these attributes.

Assumption.

The data holder can identify attributes in their private information that may also appear in external information.

Consider an instance where this assumption is incorrect; that is, the data holder misjudges which attributes are sensitive for linking. In this case, the released data may be less anonymous than what was required, and as a result, individuals may be more easily identified. In section 7, I discuss this risk and the fact that it cannot be perfectly resolved by the data holder because the data holder cannot always know what each recipient of the data knows. Further, the data holder may find it necessary to release compromising releases that are only partially anonymous. In these cases, I pose solutions that reside in policies, laws and contracts. In the remainder of this section and the next, I assume a proper quasi-identifier has been recognized.

Definition. explicit-identifier

Let T(A_1,…, A_n) be a person-specific table and Q_T(A_i,…,A_j) be a quasi-identifier for T. Further, let {A_x,…,A_y} $\subseteq Q_T$ and D be the set of direct communication methods, such as email, telephone, postal mail, etc., where with no additional information, $g_d \in$ D is a relation from T[A_x,…, A_y] to

the population reachable by g_d's communication method. Let X(s) be a random variable on the sample space s={$|g_d(\ t[A_x,\ldots,\ A_y])|$: $t \in \mathsf{T}$}. I say $\{A_x,\ldots,A_y\}$ is an *explicit identifier* of T if the expected value of X(s) is 1 and $1/\sigma$ of X(s)$\approx\infty$.

Basically, the definition above states that an explicit identifier is a set of attributes than can be used together with a direct communication method, and no additional information, to distinctly and reliably contact the entity that is the subject of those values for the attributes. Recognizing that such communications are not perfect, the definition implies the method should be almost perfect.

Definition. explicit-identifiers

Let $\mathsf{T}(A_1,\ldots,\ A_n)$ be an entity-specific table and $Q_T(A_i,\ldots,A_j)$ be a quasi-identifier for T. The explicit identifiers of T, written, E_T = {e_i : e_i is an explicit identifier of T}.

The definition above states that the explicit identifiers of a table is a set of attribute sets, where each member set is an explicit identifier of the table.

Lemma.

The explicit identifiers of table T is E_T if and only if the explicit identifiers of a quasi-identifier of T is E_T.

Example.

The following are examples of explicit identifiers: {*email address*}, {*name, address*}, {*name, phone number*}. The following are quasi identifiers, but are not explicit identifiers: {*name*}, {*Social Security number*}, {*phone*}, {*phone, Social Security number*}.

Given entity-specific data, an *anonymous data system* releases entity-specific data such that the identities of the entities that are the subject of the original data are protected. Such protection typically relies on a quasi-identifier for the original entity-specific data. The definition below defines a basic anonymous data system.

Definition. basic anonymous data system

A *basic anonymous data system*, **ADS_0**, is a nine-tuple (S, P, PT, *QI*, U, R, E, G, *f*), where the following conditions are satisfied:

1. S is the finite set of entities with attributes to be protected.
2. P is the finite set of possible entities. $S \subseteq P$.
3. PT is the finite multi-set of privately held information about each member of S. There exists a collection function, $f_c : S \rightarrow PT$, where $PT=\{k \bullet t_s : t_s = f_c(s)$ and $|f_c^{-1}(f_c(s))| = k, \forall s \in S\}$.
4. *QI* is the quasi-identifier of PT denoting attributes to be protected.
5. U is a finite set of possible entities and pseudo-entities. $P \subseteq U$.
6. R is the set of possible releases. Each release $RT \in R$ is a finite multi-set.
7. E is the collection of possible external information. $\forall T_{i=1,\ldots,m}$ where T_i is a collection of external information about a subset of the members of P, then $E = T_1 \times \ldots \times T_n$.
8. G is the set of possible relations from $R \rightarrow U$.

$$G = \left\{ (g_1, g_2) : g_1 \circ g_2 \text{ where } R \xrightarrow{g_1} E \xrightarrow{g_2} U \right\}$$

Given a *QI* for PT, written $QI_{PT} = A_i,\ldots,A_j$, a release $RT \in R$ where $RT = f(PT[QI])$, and a set of explicit identifiers named EI_{g2} where $g_2(g_1(RT)[EI_{g2}]) \subseteq U$, then

$g_1(RT) = \{k \bullet t_u[A_1,\ldots,A_m] : t_u[QI_{PT}] \in RT$, $t_u[EI_{g2}] \in E$ and $|t_u[QI_{PT}, EI_{g2}| = k$,

$\forall t_u \in E$, $QI_{PT} \subseteq A_1,\ldots,A_m$ and $EI_{g2} \subseteq A_1,\ldots,A_m\}$.

g_2 and g_2 are relations and g_2 is a direct communication method.

9. *f* is a disclosure control function such that $f:\{PT\} \rightarrow R$ and given a release $RT \in R$ where $RT = f(PT[QI])$, one of the following conditions must be satisfied:
   a. if $\exists\, g \in G, \exists\, t \in RT$, where $f(f_c(s)) = t$ and $g(f(f_c(s))) = s$ then $\exists\, u \in U$, such that $u \neq s$ and $g(f(f_c(s))) = u$.
   b. if $\exists\, (g_1, g_2) \in G$ where $GT = g_1(f(t_s[QI]))$, $\exists\, t_s[QI] \in RT$ and $t_s[QI, EI_{g2}] \in GT$ where $f_c(s) = t_s$ and $g_2(g_1(f(t_s[QI]))[EI_{g2}]) = s$, then $\exists\, t_u[QI, EI_{g2}] \in GT$ such that $t_s \neq t_u$ and $g_2(t_u[QI, EI_{g2}]) = s$.

c. Given $\mathsf{PT}(A_1,\ldots,A_n)$ and $\mathsf{RT}\ (A_w,\ldots,A_y)$, let $A_p,\ldots A_q = (\{A_1,\ldots,A_n\} - QI) \cap \{A_w,\ldots,A_y\}$. If $\exists\ g \in \mathsf{G}, \exists\ t_{s1}[A_p,\ldots,A_q] \in \mathsf{RT}$, where $f_c(s) = t_{s1}$ and $g(f(t_{s1}[QI])) = s$ and $t_{s1}[A_p,\ldots,A_q] \neq \phi$ and if $\exists\ t_{s2}[A_p,\ldots,A_q] \in \mathsf{PT}$ such that $f_c(s) = t_{s2}$ and $f(t_{s2}) = t_{s1}$ and $t_{s2}[A_p,\ldots,A_q] = t_{s1}[A_p,\ldots,A_q]$, then condition (a) or condition (b) above must be satisfied on t_{s1}.

The main property is property 9. It says that if f produces a release $\mathsf{RT} \in \mathsf{R}$ based on $\mathsf{PT}[QI]$, then there can not exist a function or composite of functions which can confidently associate any of the original subjects uniquely with their information in PT.

If an entity is correctly associated with a released tuple in RT, then the three conditions required in property 9 are: (1) there must be more than one such entity to which the tuple in the release could be associated; (2) there must be more than one such tuple in the release that could be associated with the subject; or, (3) the non-controlled information, if present, can not be accurate.

Properties 3, 7 and 8 describe multiset collections of information where collections of elements can occur as a member more than once.

The definition above describes what is termed a *basic anonymous data system*. The word "basic" is used and the subscript 0 attached because the definition does not allow for probabilistic linking or the temporal nature of data quality (i.e., older data can be less reliable). For anonymous data systems to be defined to include these issues requires a modification and extension to $\mathbf{ADS_0}$ and so, the naming convention reserves $\mathbf{ADS_1}$ and $\mathbf{ADS_2}$ and so on, for future enhancements.

Remark.

The level of protection provided by an $\mathbf{ADS_0}$ depends on the correctness of the selection of attributes within *QI*, on the specifics of *f* and on assertions and invariants that can be made about g_1 and g_2, $\forall(g_1, g_2) \in \mathsf{G}$. The validity of this remark stems directly from the definition of an $\mathbf{ADS_0}$.

S = {(*name*, Ann), (*name*, Abe), (*name*, Al)}
P = {(*name*, Dan), (*name*, Don), (*name*, Dave), (*name*, Ann), (*name*, Abe), (*name*, Al)}
PT(*name*, *birth date*, *ZIP*, *diagnosis*) **:**

Name	Birth date	ZIP	Diagnosis
Ann	10/2/61	02139	Cardiac
Abe	7/14/61	02139	Cancer
Al	3/8/61	02138	Liver

QI = {*name*, *birth date*, *ZIP*}
U = {(*name*, Jcd), (*name*, Jwq), (*name*, Jxy), (*name*, Dan), (*name*, Don), (*name*, Dave),
(*name*, Ann), (*name*, Abe), (*name*, Al) }
E(*name*, *birth date*, *ZIP*) **:**

Name	Birth date	ZIP
Ann	10/2/61	02139
Abe	7/14/61	02139
Al	3/8/61	02138

g_2 = a direct communication channel that operates on the *name* attribute.
G as the set of all possible relations from R to U consistent with property 8 in the definition of an **ADS$_0$**

Figure 19 Values for S, P, PT, *QI*, U and E

In the following examples, I assume the values for S, P, PT, *QI*, U, and E shown in Figure 19. These values are consistent with the presentations in Figure 15, Figure 16 and Figure 17.

Example (identity release).

Given the assignments in Figure 19, and the following definition for *f* that constructs RT as a copy of PT, the system **A**(S, P, PT, *QI*, U, {RT}, E, G, *f*) is not an **ADS$_0$**.

f is defined as follows:

step 1. Let RT be ∅

step 2. $\forall t \in$ PT, RT $\leftarrow$ RT $\cup$ {*t*}

Proof:

Let g_1 be the relation g_1(*name*, *birth date*, *ZIP*, *diagnosis*) on RT.

Therefore **A** is insecure and a disclosure is made, so **A** is not an **ADS$_0$**.

Example (complete suppression).

Given the definitions in Figure 19, and the following definition for *f* that constructs RT as a blank table, the system **A**(S, P, PT, *QI*, U, {RT}, E, G, *f*) is an **ADS$_0$**.

f is defined as follows:

step 1. Let RT be ∅

step 2. $\forall t \in$ PT, RT ← RT ∪ {null, null, null, null}

Note. RT is a multi-set, so duplicates are maintained.

Proof:

The first two conditions of property 9 in the definition of an **ADS$_0$** are both satisfied $\forall t \in$ RT.

Therefore **A** is considered secure, so **A** is an **ADS$_0$**.

The two examples above demonstrate the natural tension that exists in disclosure control. At one end is specificity and usefulness, which is not secure, and at the other end is distortion and security, which is not useful. These opposites pose a continuum of disclosure control options along which tradeoffs must be made. I used an information theoretic (entropy) metric and measured the distortion to data caused by common disclosure limitation techniques (see Figure 14 on page 24) and then plotted the measures along the continuum. The relative ordering of the results is shown below Figure 20.

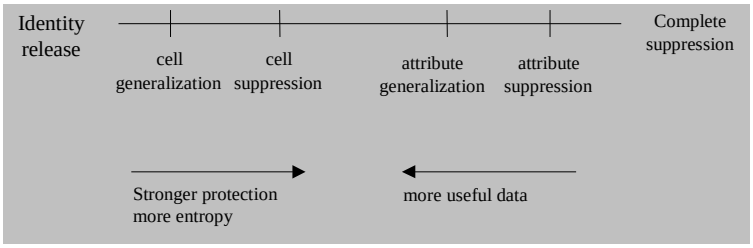


Figure 20 Relative comparison of techniques

The technique cell generalization is generalization enforced at the cell level and likewise cell suppression is suppression enforced at the cell level. Similarly, attribute generalization is generalization enforced at the attribute level and attribute suppression is suppression enforced at the attribute level. Do not interpret the tick marks along the continuum as points. Each of these techniques had results in a range along the continuum and the ranges overlapped; further there was significant variation depending on the

character of the data. However, the tick marks do provide a relative ordering of the medians of average case results.

I now present three protection models for **ADS$_0$**. These are *wrong-map*, *null-map* and *k-map* as defined below.

Definition. null-map protection

Let **A** be an **ADS$_0$**, $f(\mathsf{PT}) = \mathsf{RT}$ and $\mathsf{R} \in \mathsf{RT}$. If $\forall t \in \mathsf{RT}$, there does not exist $g \in \mathsf{G}$ where $g(t) \in \mathsf{S}$, then **A** adheres to *null map protection*.

In *null-map protection* each tuple in the released information may or may not map to an actual entity in the population P, but none of the tuples can be mapped to an entity in the set of subjects S. Examples of disclosure limitation techniques that can achieve null-map protection include strong encryption of the *QI*, extensive swapping of the values in *QI* and systematic use of additive noise. Figure 16 provides an example.

Definition (wrong-map protection).

Let **A** be an **ADS$_0$**, $f(\mathsf{PT}) = \mathsf{RT}$ and $\mathsf{R} \in \mathsf{RT}$. If $|\mathsf{RT}| > 2$ and $\forall t \in \mathsf{RT}$, $\exists g \in \mathsf{G}$ where $f(f_c(s)) = t$, and $g(f(f_c(s))) = s$ and there does not exist $g' \in \mathsf{G}$ where $g' \neq g$ such that $g'(t) \in \mathsf{S}$, then **A** adheres to *wrong map protection.*

Wrong map protection requires each tuple in the released information to be identified to only one entity in subjects but that entity is not the entity to which the original information was collected. The **ADS$_0$** requirement ensures the values with attributes outside *QI* contained in the release are not the same as those originally collected. Notice if there exists only one entity in the subjects S, then wrong-map protection cannot be done and with only two entities in S, the release is compromised. An example of a disclosure limitation technique that can achieve wrong map protection is swapping the attributes of *QI* as a unit.

Definition (*k*-map protection).

> Let **A** be an **ADS$_0$**, $f(\mathsf{PT}) = \mathsf{RT}$ and $\mathsf{R} \in \mathsf{RT}$. If $\forall t \in \mathsf{RT}$, $\exists\, g \in \mathsf{G}$, where $f(f_c(s)) = t$ and $g(f(f_c(s))) = s$ and $\{u_1, u_2, u_{k-1}\} \in \mathsf{U}$ such that for i=1,…, k-1, $u_i \neq s$, and $g(f(f_c(s))) = u_i$, then **A** adheres to *k-map protection*.

k-map protection maintains the invariant that each tuple in the released information refers indistinctly to at least *k* members of U. Notice that *k* does not rely on $|\mathsf{S}| > k$ or on $|\mathsf{RT}| > k$. Figure 17 provides an example.

The protection models *k*-map, null-map and wrong-map provide a means for characterizing the kind of protection provided to a release of information. Of course a release may be anonymous, but proving it in the absence of a protection model is extremely difficult. Optimal releases that offer adequate protection with minimal distortion are believed to typically require a combination of disclosure limitation techniques and a combination of protection models.

6 Overview of four disclosure control systems

In this section I examine real-world computational systems that attempt to produce anonymous data, but before proceeding here is a quick review of what has been covered so far. Section 1 motivated this work based on the increased demands for sharing person-specific information enabled by today's technological setting. Section 2 provided real-world examples of why the problem of producing anonymous information is so difficult. Section 3 reviewed related work and noted that much of the prior work on inference control appears inadequate for today's real-world use. Section 4 surveyed disclosure limitation techniques and provided a framework for reasoning about disclosure control. Finally, section 5, formally introduced a basic anonymous data system, **ADS$_0$**, and introduced three protection models that can be used to determine adequate protection for an **ADS$_0$**.

I now present four computational systems that attempt to maintain privacy while releasing electronic information. These systems are: (1) my Scrub System, which locates personally-identifying information in letters between doctors and notes written by clinicians; (2) my Datafly II System, which generalizes and suppresses values in field-structured data sets; (3) Statistics Netherlands' μ-Argus System, which is becoming a European standard for producing public-use data; and, (4) my *k*-Similar algorithm, which produces optimal results in comparison to Datafly and μ-Argus. I assess the anonymity protection provided by each of these systems in terms of whether each system is an **ADS$_0$**. This presentation returns to an informal style.

6.1 The Scrub System

My Scrub System locates and replaces personally identifying information in text documents and in textual fields of the database. Scrub and Scrub-like systems have been used to automatically gather person-specific information directly from textual documents found on the World Wide Web. A close examination of two different computer-based patient record systems, Boston's Children's Hospital [34] and Massachusetts General Hospital [35], quickly revealed that much of the medical content resided in the letters between physicians and in the shorthand notes of clinicians. This is where providers discussed findings, explained current treatment and furnished an overall view of the medical condition of the patient.

At present, most institutions have few releases of medical data that include these notes and letters, but new uses for this information is increasing; therefore, the desire to release this text is also increasing. After all, these letters and notes are a valuable research tool and can corroborate the rest of the record. The fields containing the diagnosis, procedure and medication codes when examined alone can be incorrect or misleading. A prominent physician stated at a recent conference that he purposefully places incorrect codes in the diagnosis and procedure fields when such codes would reveal sensitive information about the patient [36]. Similarly, the diagnosis and procedure codes may be up-coded for billing purposes. The General Accounting Office estimates that as much as 10% of annual Federal health care expenditures, including Medicare, are lost to fraudulent provider claims [37]. If these practices become widespread, they will render the administrative medical record useless for clinical research and may already be problematic for retrospective investigation. Clinical notes and letters may prove to be the only reliable artifacts.

The Scrub System provides a methodology for removing personally identifying information in medical writings so that the integrity of the medical information remains intact even though the identity of the patient remains confidential. This process is termed *scrubbing*. Protecting patient confidentiality in raw text is not as simple as searching for the patient's name and replacing all occurrences with a pseudo name. References to the patient are often quite obscure; consider for example:

> "…he developed Hodgkins while acting as the U.S. Ambassador to England and was diagnosed by Dr. Frank at Brigham's."

Clinicians write text with little regard to word-choice and in many cases without concern to grammar or spelling. While the resulting "unrestricted text" is valuable to understanding the medical condition and

treatment of the patient, it poses tremendous difficulty to scrubbing since the text often includes names of other care-takers, family members, employers and nick names.

I examined electronically stored letters written by clinical specialists to the physician who referred the patient. The letter in Figure 21 is a fictitious example modeled after those studied. It contains the name and address of the referring physician, a typing mistake in the salutation line, the patient's nick name, and references to another care-taker, the patient's athletic team, the patient's mother and her mother's employer and phone number. Actual letters are often several pages in length.

Wednesday, February 2, 1994

Marjorie Long, M.D.
St. John's Hospital
Huntington 18
Boston, MA 02151

RE: Virginia Townsend
CH#32-841-09787
DOB 05/26/86

Dear Dr. Lang:

I feel much better after seeing Virginia this time. As you know, Dot is a 7 and 6/12 year old female in follow up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. Frank at Brigham's. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the U. S. Junior Gymnastics team. We will contact Mrs. Hodgkins in a week at Marina Corp 473-1214 to schedule a follow-up visit for her daughter.

Patrick Hayes, M.D. 34764

Figure 21. Sample letter reporting back to a referring physician.

February, 1994

Erisa Cosborn, M.D.
Brighaul Hospital
Alberdam Way
Peabon, MA *02100*

RE: *Kathel Wallams*
CH#*18-512-32871*
DOB *05/86*

Dear Dr. *Jandel*:

I feel much better after seeing *Kathel* this time. As You know, *Cob* is a 7 and 6/12 year old female in follow-up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. *Wandel* at *Namingham*'s. She is currently on Lily Human Insulin and is growing and Gaining weight normally. She will start competing again with the . We will Contact Mrs. *Learl* in a week at *Garlaw* Corp *912-8205* to schedule a follow-up visit for her daughter.

Mank Brones, M.D. *21075*

Figure 22. **Scrub System applied to sample in Figure 21.**

Figure 21 shows a sample letter and Figure 22 shows its scrubbed result. Notice in the scrubbed result (Figure 22) that the name of the medication remained but the mother's last name was correctly replaced. Dates were changed to report only month and year. The reference "U.S. Junior Gymnastics team" was suppressed since Scrub was not sure how to replace it. The traditional approach to scrubbing is straightforward search and replace, which misses these references; this is shown in Figure 23.

Wednesday, February 2, 1994

Marjorie Long, M.D.
St. John's Hospital
Huntington 18
Boston, MA 02151

RE: *Kathel Wallams*
CH#*18-512-32871*
DOB 05/26/86

Dear Dr. Lang:

I feel much better after seeing *Kathel* this time. As you know, Dot is a 7 and 6/12 year old female in follow up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. Frank at Brigham's. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the U. S. Junior Gymnastics team. We will contact Mrs. Hodgkins in a week at Marina Corp 473-1214 to schedule a follow-up visit for her daughter.

Mank Brones, M.D. *21075*

Figure 23. **Search-and Replace applied to sample in Figure 1-8.**

The Scrub System was modeled after a human approach to the problem. It uses templates and localized knowledge to recognize personally identifying information. In fact, the work on Scrub shows

that the recognition of personally identifying information is strongly linked to the common recording practices of society. For example, Fred and Bill are common first names and Miller and Jones are common last names; knowing these facts makes it easier to recognize them as likely names. Common facts, along with their accompanying templates of use, are considered commonsense knowledge; the itemization and use of commonsense knowledge is the backbone of Scrub.

The Scrub System utilizes numerous detection algorithms competing in parallel to label contiguous characters of text as being a proper name, an address block, a phone number, and so forth. Each detection algorithm recognizes a specific kind of information, where recognizable kinds of information can be thought of as fields such as *first name*, *last name*, *street address*, and *date*. There is at least one detection algorithm for each kind of information.

Detection algorithms in Scrub use local knowledge sources, such as lists of area codes and first names and helping routines such as those that determine whether words "sound" like medical terms or last names. Each algorithm tries to identify occurrences of its assigned field of information.

The Scrub System accurately found 99-100% of all personally identifying references in more than 3,000 letters between physicians, while the straightforward approach of global search-and-replace properly located no more than 30-60% of all such references; these values are summarized in Figure 24. The higher figure for search and replace includes using additional information stored in the database to help identify the attending physician's name, identifying number and other information. Since the letters were properly formatted, the heading block was easily detected and compositional cues were available using keywords like "Dear." This dramatically improved the results of the search-and-replace method to around 84%; however, most references to family members, additional phone numbers, nick names and references to the physician receiving the letter were still not detected, whereas Scrub was able to correctly identify and replace these instances.

Method	**Letters**
Straight search	37%
Search with cues	84%
Scrub(threshold 0.8)	98%
Scrub(threshold 0.7, false positive reduction)	100%

Figure 24 Comparisons of Scrub to standard techniques

Despite this apparent success, the Scrub System merely de-identifies information and cannot guarantee anonymity. Even though all explicit identifiers such as name, address and phone number are

removed or replaced, it may be possible to infer the identify of an individual. Consider the text in Figure 25.

> "At the age of two she was sexually assaulted. At the age of three she set fire to her home. At the age of four her parents divorced. At the age of five she was placed in foster care after stabbing her nursery school teacher with scissors."

Figure 25 Sample de-identified text

If her life continues to progress in this manner, by the age of eight she may be in the news, but nothing in this text required scrubbing even though there would probably exist only one such child with this history. An overall sequence of events can provide a preponderance of details that identify an individual. This is often the case in mental health data and discharge notes.

Scrub as an anonymous data system

Scrub uses the following disclosure limitation techniques: de-identification, equivalence class substitution, generalization, and suppression. Below is a description of the framework in which Scrub operates.

S = {*subjects whose information is discussed in textual documents PT*}

P = *set of all people whose information could possibly be PT*

PT = *set of documents about S*

QI = set of attributes for which Scrub detectors are available

$U = \{d_1 \times \ldots \times d_n\} \cup P$

RT = *Scrub*(PT)

E = set of publicly available information in today's society

G = set of standard communication methods.

f = Scrub System

The system **A**(S, P, PT, *QI*, U, {RT}, E, G, *Scrub*) is <u>not</u> an $\mathbf{ADS_0}$.

Informal proof.

Assume **A** is an $\mathbf{ADS_0}$.

Let p_i be the person who is the subject of the text in Figure 25.

E includes newspaper reports and phone books that include p_i's family.

By simply linking the information, as was demonstrated in Figure 12, p_i can be re-identified, violating property 9 of an **ADS$_0$**.

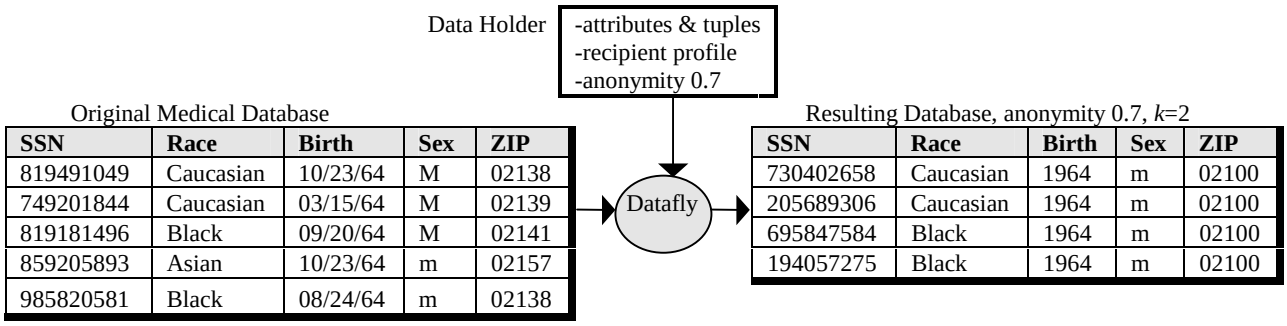
So, **A** is not an **ADS$_0$**.

Although Scrub reliably locates explicitly identifying information in textual documents, it merely de-identifies the result because its detectors are aimed primarily at explicitly identifying values. In my earlier examples, such as the voter list example in section 2 on page 18, I showed in field-structured databases that de-identification typically provides insufficient protection. Other values remaining in the data can combine uniquely to identify subjects. The Scrub work demonstrates that this is as true in textual documents as it is in field-structured databases. But perhaps more importantly, the Scrub work implies that solving the problem in one data format (either textual documents or field-structured databases) will reveal comparable strategies for solving the problem in the other format. In the next subsections, I present some proposed solutions for solving the problem in field-structured databases.

The Scrub System is both troublesome and insightful in another regard. While Scrub is inadequate for privacy protection, it is quite useful in automatically detecting and gathering personally identifying information from email messages, World Wide Web pages, and other textual information appearing in an electronic format and then using the results to draw damaging inferences from other publicly available field-structured data sets. In this way, Scrub demonstrates the symbiotic relationship between data detective tools and data protection tools. Re-identification experiments and the tools used to accomplish re-identifications improve our understanding of the identifiability of data and our tools for rendering data sufficiently anonymous.

6.2 The Datafly II System

My Datafly and Datafly II Systems provide the most general information useful to the recipient in field-structured databases. From now on, the term Datafly will refer to the Datafly II System unless otherwise noted. Datafly maintains anonymity in released data by automatically substituting, generalizing and suppressing information as appropriate. Decisions are made at the attribute and tuple level at the time of database access, so the approach can be incorporated into role-based security within an institution as well as in exporting schemes for data leaving an institution. The end result is a subset of the original database that provides minimal linking and matching of data because each tuple matches as many people as the data holder specifies.



SSN	Race	Birth	Sex	ZIP
819491049	Caucasian	10/23/64	M	02138
749201844	Caucasian	03/15/64	M	02139
819181496	Black	09/20/64	M	02141
859205893	Asian	10/23/64	m	02157
985820581	Black	08/24/64	m	02138

SSN	Race	Birth	Sex	ZIP
730402658	Caucasian	1964	m	02100
205689306	Caucasian	1964	m	02100
695847584	Black	1964	m	02100
194057275	Black	1964	m	02100

Figure 26. **User-level overview of the Datafly System**

Figure 26 provides an overview of the Datafly System from the data holder's perspective for generating a table for release. The original table is shown on the left. Input to the Datafly System is the original privately held table and some specifications provided by the data holder. Output is a table whose attributes and tuples correspond to the anonymity level specified by the data holder; in Figure 26 the anonymity level is noted as being 0.7. These terms and the process used by Datafly to generate a table for release are discussed in the following paragraphs.

Before any releases are generated, each attribute in the original table is tagged as using either an equivalence class substitution algorithm or a generalization routine when its associated values are to be released. If values of an attribute tagged as using equivalence class substitution are to be released, made-up alternatives replace values of the attribute in the released data. The Social Security number attribute labeled *SSN* provides an example in Figure 26.

Alternatively, if an attribute is tagged as using generalization, then an accompanying generalization hierarchy is assigned to the attribute; example hierarchies are shown in Figure 18 on page 28. The Datafly System iteratively computes increasingly less specific versions of the values for the attribute until eventually the desired anonymity level is attained. For example, the *birth date* attribute would first have the full month, day and year for each value. If further generalization were necessary, only the month and year would be used, and then only the year and so on, as the values get less and less specific, moving up the generalization hierarchy. The iterative process ends when there exists k tuples having the same values assigned across a group of attributes (or quasi-identifier); this is termed a k requirement and provides the basis for k-map protection. [Note in the earliest version of Datafly, k was enforced on each attribute individually and a complicated requirement was enforced across attributes; but in later versions which are named Datafly II, k is enforced across the quasi-identifier as described here.] In Figure 26 the quasi-identifier under consideration, because of the size of the database shown, is only {*Birth*, *Sex*, *ZIP*} and k=2; therefore, in the released data, there are at least two tuples for each combination of {*Birth*, *Sex*, *ZIP*} released. If *Race* had been included in the quasi-identifier, then the

values for the entire attribute would have been generalized to the singleton value of "Person" in order to achieve the anonymity requirement of *k*=2.

To use the system, the data holder (1) declares specific attributes and tuples in the original private table as being eligible for release. The data holder also (2) provides a list in which a number from 0 to 1 is assigned to each attribute eligible for release denoting the amount of distortion that can be tolerated by the recipient; a 0 value means minimal distortion and a value of 1 indicates maximal distortion. I term such a list a *profile*. The data holder (3) groups a subset of the released attributes into one or more quasi-identifiers and provides a second profile that identifies the likelihood each attribute within a quasi-identifier will be used for linking; a 0 value means not likely and a value of 1 means highly probable. Finally, the data holder (4) specifies a minimum overall anonymity level that computes to a value of *k* and (5) a threshold (called *maxDrop*) that determines the maximum number of tuples that can be suppressed.

Datafly then produces the released table from the eligible attributes and tuples of the private table such that each value of a quasi-identifier in the released table appears in at least *k* tuples. The *k* requirement is accomplished by generalizing attributes within a quasi-identifier as needed and suppressing no more than *maxDrop* tuples.

In Figure 26, notice how the record containing the Asian entry was removed; Social Security numbers were automatically replaced with made-up alternatives; birth dates were generalized to the year and ZIP codes to the first three digits. In the next two paragraphs I examine the overall anonymity level and its relationship to *k*.

The overall anonymity level is a number between 0 and 1 that relates to the minimum *k* for each quasi-identifier. An anonymity level of 0 provides the original data and a level of 1 forces Datafly to produce the most general data possible given the profile of the recipient. All other values of the overall anonymity level between 0 and 1 determine the operational value for *k*. (The institution is responsible for mapping the anonymity level to particular values of *k* though we can provide some guidelines.) Information within each attribute is generalized as needed to attain the minimum *k* and *outliers*, which are extreme values not typical of the rest of the data, may be removed. Upon examination, the resulting data, every value assigned to each quasi-identifier will occur at least *k* times with the exception of one-to-one replacement values, as is the case with Social Security numbers.

Anonymity (*A*)	*k*	Birth Date	maxDrop%
1			
--- .9 ---	493	24	4%
--- .8 ---	438	24	2%
--- .7 ---	383	12	8%
--- .6 ---	328	12	5%
--- .5 ---	274	12	4%
--- .4 ---	219	12	3%
--- .3 ---	164	6	5%
--- .2 ---	109	4	5%
--- .1 ---	54	2	5%
0			

Figure 27. Anonymity generalizations for Cambridge voters' data with corresponding values of *k*.

Figure 27 shows the relationship between *k* and selected anonymity levels (*A*) using the Cambridge voters' database. As *A* increased, the minimum requirement for *k* increased, and in order to achieve the *k*-based requirement, values within an attribute in a quasi-identifier, for example, *Birth Date*, were re-coded as shown. Outliers were excluded from the released data, and their corresponding percentages of *N* (where *N* is the number of tuples in the privately held table eligible for release) are noted. An anonymity level of 0.7, for example, required at least 383 occurrences of every value of the quasi-identifier. To accomplish this in only *Birth Date*, for example, required re-coding dates to reflect only the birth year. Even after generalizing over a 12 month window, the values of 8% of the voters still did not meet the requirement so these voters were dropped from the released data.

In addition to an overall anonymity level, the data holder also provides a profile of the needs of the person who is to receive the data by specifying for each attribute that is to be in the release whether the recipient could have or would use information external to the database that includes data within that attribute. That is, the data holder estimates on which attributes the recipient might link outside knowledge. Thus, each attribute has associated with it a profile value between 0 and 1, where 0 represents full trust of the recipient or no concern over the sensitivity of the information within the attribute, and 1 represents full distrust of the recipient or maximum concern over the sensitivity of the attribute's contents. Semantically related attributes that are sensitive to linking, with the exception of one-to-one replacement attributes, are treated as a single concatenated attribute (a quasi-identifier) that must meet the minimum *k* requirement, thereby thwarting linking attempts that use combinations of attributes. The role of these profiles is to help select which attribute within the quasi-identifier will be selected for generalization. If all attributes in the quasi-identifier have the same value, then the attribute having the greatest number of distinct values will be generalized.

Consider the profiles of a doctor caring for a patient, a clinical researcher studying risk factors for heart disease, and a health economist assessing the admitting patterns of physicians. Clearly, these

profiles are all different. Their selection and specificity of attributes are different; their sources of outside information on which they could link are different; and their uses for the data are different. From publicly available birth certificates, driver license, and local census databases, the birth dates, ZIP codes and gender of individuals are commonly available along with their corresponding names and addresses; so these attributes could easily be used for re-identification. Depending on the recipient, other attributes may be even more useful. If the recipient is the patient's caretaker within the institution, the patient has agreed to release this information to the care-taker, so the profile for these attributes should be set to 0 to give the patient's caretaker full access to the original information.

When researchers and administrators make requests that require less specific information than that originally provided within sensitive attributes, the corresponding profile values should warrant a number as close to 1 as possible, but not so much so that the resulting generalizations provide useless data to the recipient. But researchers or administrators bound by contractual and legal constraints that prohibit their linking of the data are trusted, so if they make a request that includes sensitive attributes, the profile values would ensure that each sensitive attribute adheres only to the minimum *k* requirement.

The goal is to provide the most general data that are acceptably specific to the recipient. Since the profile values are set independently for each attribute, particular attributes that are important to the recipient can result in less generalization than other requested attributes in an attempt to maintain the usefulness of the data. A profile for data being released for public use, however, should be 1 for all sensitive attributes to ensure maximum protection. The purpose of the profiles are to quantify the specificity required in each attribute and to identify attributes that are candidates for linking; and in so doing, the profiles identify the associated risk to patient confidentiality for each release of data.

Numerous tests were conducted using the Datafly System to access a pediatric medical record system. Datafly processed all queries to the database over a spectrum of recipient profiles and anonymity levels to show that all attributes in medical records can be meaningfully generalized as needed since any attribute can be a candidate for linking. Of course, which attributes are most important to protect depends on the recipient. Diagnosis codes have generalizations using the International Classification of Disease (ICD-9) hierarchy (or other useful semantic groupings). Geographic replacements for states or ZIP codes generalize to use regions and population size. Continuous variables, such as dollar amounts and clinical measurements, can be converted to discrete values; however, their replacements must be based on meaningful ranges in which to classify the values; of course this is only done in cases where generalizing these attributes is necessary.

In the real-world example mentioned earlier on page 19, the Group Insurance Commission in Massachusetts (GIC) collected patient-specific data with almost 100 attributes of information per

physician visit by more than 135,000 state employees, their families and retirees. In a public hearing, GIC reported giving a copy of the data to a researcher, who in turn stated that she did not need the full date of birth, just the birth year [38]. The average value of *k* based only on {*birth date*, *gender*} for that population is 3, but had the researcher received only {*year of birth*, *gender*}, the average value of *k* would have increased to 1125. Furnishing the most general information the recipient can use minimizes unnecessary risk to patient confidentiality.

Datafly as an anonymous data system

Datafly uses the following disclosure limitation techniques: de-identification, equivalence class substitution, generalization, and suppression. Below is a description of the framework in which Datafly operates.

S = {*subjects whose information is included in PT*}

P = *set of all people whose information could possibly be in PT*

PT = *privately held information about S*

QI = set of attributes with replications in E

U = {*existence of people implied by equivalence class assignments*} $\cup P$

RT = *Datafly*(PT)

E = set of publicly available information in today's society

G = set of standard communication methods.

f = Datafly System

The system **A**(S, P, PT, *QI*, U, {RT}, E, G, *Datafly*) is an $\mathbf{ADS_0}$.

Informal proof.

If *QI* contains all attributes replicated in E, **A** adheres to *k*-map protection,

where *k* is enforced on RT. That is, for each value of *QI* released in RT,

there are at least *k* tuples having that value.

So, **A** is an $\mathbf{ADS_0}$.

Datafly and Scrub use the same disclosure limitation techniques even though they operate on different kinds of data. But unlike Scrub, Datafly is an $\mathbf{ADS_0}$ in cases where the quasi-identifier is correctly chosen because in those cases each tuple released by Datafly will indistinctly map to at least *k* entities. Scrub provides no such protection.

6.3 The μ-Argus System

In 1996, The European Union began funding an effort that involves statistical offices and universities from the Netherlands, Italy and the United Kingdom. The main objective of this project is to develop specialized software for disclosing public-use data such that the identity of any individual contained in the released data cannot be recognized. Statistics Netherlands has already produced a first version of a program named μ-Argus that seeks to accomplish this goal [39]. The μ-Argus program is considered by many as the official confidentiality software of the European community. A presentation of the concepts on which μ-Argus is based can be found in Willenborg and De Waal [40].

The program μ-Argus, like the Datafly System, provides protection by enforcing a *k* requirement on the values found in a quasi-identifier. It generalizes values within attributes as needed, and removes extreme outlier information from the released data. The user provides a value of *k* and specifies which attributes are sensitive by assigning a value to each attribute between 0 and 3 denoting "not identifying," "identifying," "more identifying," and "most identifying," respectively. The program then identifies rare and therefore unsafe combinations by testing 2- or 3-combinations across attributes declared to be identifying. Unsafe combinations are eliminated by generalizing attributes within the combination and by local cell suppression. Rather than removing entire tuples when one or more attributes contain outlier information as is done in the Datafly System, the μ-Argus System simply suppresses or blanks out the outlier values at the cell-level. This process is called *cell suppression* [39]. The resulting data typically contain all the tuples and attributes of the original data, though values may be missing in some cell locations.

In Figure 28 there are many Caucasians and many females, but only one female Caucasian in the database. Figure 29 shows the results from applying the Datafly system to the data provided in Figure 28. The given profile identifies only the demographic attributes as being likely for linking and $k = 2$. The data are being made available for semi-public use so the Caucasian female tuple was dropped as an outlier.

Figure 30 shows the results from applying the approach of the μ-Argus system with $k = 2$ to the data in Figure 28. SSN was marked as being "most identifying," the birth, sex, and ZIP attributes were marked as being "more identifying," and the ethnicity attribute was simply marked as "identifying." Combinations across these were examined; the resulting suppressions are shown. The uniqueness of the Caucasian female is suppressed; but, there still remains a unique tuple for the Caucasian male born in

1964 who lives in the 02138 ZIP code. I will now step through how the μ-Argus program produced the results in Figure 30.

The first step is to check that each identifying attribute adheres to *k* requirement. Then, *pairwise* combinations are examined for each pair that contains the "most identifying" attribute (in this case, SSN) and those that contain the "more identifying" attributes (in this case, birth date, sex and ZIP). Finally, 3-combinations are examined that include the "most" and "more" identifying attributes. Obviously, there are many possible ways to rate these identifying attributes and, unfortunately, different identification ratings yield different results. The ratings presented in this example produced the most secure result using the μ-Argus program, though admittedly one may argue that too many specifics remain in the data for it to be released for public use.

Each unique combination of values found within sensitive attributes constitutes a bin. When the number of occurrences of such a combination is less than the minimum required bin size, the combination is considered unique and termed an outlier. Clearly for all combinations that include the SSN, all such combinations are unique. One value of each outlier combination must be suppressed. For optimal results, the μ-Argus program suppresses values that occur in multiple outliers where precedence is given to the value occurring most often. The final result is shown in Figure 30. The responsibility of when to generalize and when to suppress resides with the user. For this reason, the μ-Argus program operates in an interactive mode so the user can see the effect of generalizing and can then select to undo the step.

SSN	Ethnicity	Birth	Sex	ZIP	Problem
819181496	Black	09/20/65	m	02141	shortness of breath
195925972	Black	02/14/65	m	02141	chest pain
902750852	Black	10/23/65	f	02138	hypertension
985820581	Black	08/24/65	f	02138	hypertension
209559459	Black	11/07/64	f	02138	obesity
679392975	Black	12/01/64	f	02138	chest pain
819491049	Caucasian	10/23/64	m	02138	chest pain
749201844	Caucasian	03/15/65	f	02139	hypertension
985302952	Caucasian	08/13/64	m	02139	obesity
874593560	Caucasian	05/05/64	m	02139	shortness of breath
703872052	Caucasian	02/13/67	m	02138	chest pain
963963603	Caucasian	03/21/67	m	02138	chest pain

Figure 28. Sample database.

SSN	Ethnicity	Birth	Sex	ZIP	Problem
902387250	Black	1965	m	02140	shortness of breath
197150725	Black	1965	m	02140	chest pain
486062381	Black	1965	f	02130	hypertension
235978021	Black	1965	f	02130	hypertension
214684616	Black	1964	f	02130	obesity
135434342	Black	1964	f	02130	chest pain
458762056	Caucasian	1964	m	02130	chest pain
860424429	Caucasian	1964	m	02130	obesity
259003630	Caucasian	1964	m	02130	shortness of breath
410968224	Caucasian	1967	m	02130	chest pain
664545451	Caucasian	1967	m	02130	chest pain

Figure 29. Results from applying the Datafly System to the data in Figure 28.

SSN	Ethnicity	Birth	Sex	ZIP	Problem
	Black	1965	m	02141	shortness of breath
	Black	1965	m	02141	chest pain
	Black	1965	f	02138	hypertension
	Black	1965	f	02138	hypertension
	Black	1964	f	02138	obesity
	Black	1964	f	02138	chest pain
	Caucasian	1964	m	02138	chest pain
			f	*02139*	hypertension
	Caucasian	1964	m	02139	obesity
	Caucasian	1964	m	02139	shortness of breath
	Caucasian	1967	m	02138	chest pain
	Caucasian	1967	m	02138	chest pain

Figure 30. Results from applying the μ-Argus system approach to the data in Figure 28.

I will briefly compare the results of these two systems. In the Datafly System, generalizing across a quasi-identifier ensures that the corresponding tuples will adhere to the *k* requirement. This is demonstrated in Figure 29. The μ-Argus program however, only checks 2 or 3 combinations; there may exist unique combinations across 4 or more attributes that would not be detected. For example, Figure 30 still contains a unique tuple for a Caucasian male born in 1964 that lives in the 02138 ZIP code, since there are 4 characteristics that combine to make this tuple unique, not 2. Treating a quasi-identifier as a single attribute that must adhere to the *k* requirement, as done in the Datafly System provides more secure releases of data. Further, since the number of attributes, especially demographic attributes, in a medical database is large, this may prove to be a serious handicap when using the μ-Argus system with medical data.

μ-Argus as an anonymous data system

μ-Argus uses the following disclosure limitation techniques: de-identification, generalization, and suppression. Below is a description of the framework in which μ-Argus operates.

S = {*subjects whose information is included in PT*}

P = *set of all people whose information could possibly be in PT*

PT = *privately held information about S*

QI = set of attributes with replications in E

U = P

RT = *μ-Argus* (PT)

E = set of publicly available information in today's society

G = set of standard communication methods.

f = μ-Argus System

The system **A**(S, P, PT, *QI*, U, {RT}, E, G, *μ-Argus*) is not an $\mathbf{ADS_0}$.

Informal proof.

Let PT = data in Figure 28.

There can exist fewer than *k* tuples in RT having the same values across *QI*,

as shown in Figure 30.

So, *k*-map protection is not provided and **A** is not an $\mathbf{ADS_0}$.

6.4 The *k*-Similar Algorithm

My more recent work examines all combinations of values within sensitive attributes and releases an optimal solution with respect to minimal distortion due to generalization and suppression. A metric is introduced that shows Datafly can over distort data. The basis of the metric relies on combining generalization and suppression of an attribute into a common hierarchy of values and then measuring distortion as the distance of between the original value and the released value in the value generalization hierarchy (VGH) for the attribute. Figure 31 shows the value generalization hierarchy for the *ZIP* codes (02138, 02139, 02141, 02142, *****). The suppressed value, *****, is the maximal singleton value of the VGH denoting that the next action after generalizing a value to its maximum of 021** is to suppress the value altogether. The domains, which are found at each level of the VGH,

themselves form a partial ordering that I term the domain generalization hierarchy (DGH). The distance measurement between a value in the ground domain (Z_0) and it distorted value (*d*) found in the VGH for Z_0 (written VGH_{Z0}) is computed as the number of levels up VGH_{Z0} that *d* appears. This value can be computed for an entire table by summing the distances of all values in the released table. This is reflected in the formula that appears in Figure 31.

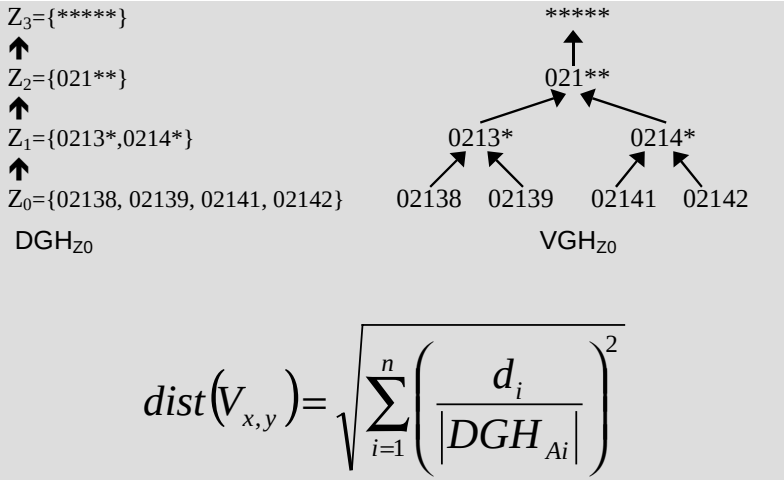


Figure 31 Distance measurement using generalization and suppression

The overall distance measurement for a released table is a measure of distortion for the released table. *Precision* then, can be computed as one less this result. This expression is provided in

$$prec(RT) = 1 - \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N} \frac{h}{|DGH_{Ai}|}}{|RT| \bullet |N_A|}$$

Figure 32 Precision metric for a released table

My *k*-Similar algorithm uses an extension of the distance metric in Figure 31 to grow clusters of tuples such that the tuples of a cluster have values of a quasi-identifier that are "closest neighbors" based on the distance of their values to a common ancestor value in a VGH. The clusters grow until each cluster has at least *k* tuples. (This is in contrast to the well-known *k*-cluster algorithm in which the tuples are partitioned into *k* clusters. In *k*-Similar, there may by one or more clusters, not necessarily *k* clusters, but all clusters will have at least *k* members.) In the end, each cluster adheres to the *k* requirement and

performing the implied cell generalizations and suppressions to make the values the same guarantees that the overall result adheres to *k*-map protection. Data released from *k*-Similar is an optimal solution in that the data are minimally distorted while still providing adequate protection. It uses generalization and suppression performed at the cell level to deliver optimal results.

As in Datafly, a profile stating the recipient's preferences is provided so that decisions among values that are equally and minimally distorting is based on recipient preference; this maintains the overall usefulness of the data.

SSN	Ethnicity	Birth	Sex	ZIP	Problem
486753948	Black	1965	m	02141	short of breath
758743753	Black	1965	m	02141	chest pain
976483662		1965	f	0213*	hypertension
845796834		1965	f	0213*	hypertension
497306730	Black	1964	f	02138	obesity
730768597	Black	1964	f	02138	chest pain
348993639	Caucasian	1964	m	0213*	chest pain
459734637		1965	f	0213*	hypertension
385692728	Caucasian	1964	m	0213*	obesity
537387873	Caucasian	1964	m	0213*	short of breath
385346532	Caucasian	1967	m	02138	chest pain
349863628	Caucasian	1967	m	02138	chest pain

Figure 33 Results from applying *k*-Similar to data in Figure 28

Figure 33 shows the results from *k*-Similar using the data in Figure 28 where *k*=2. *SSN* values have been replaced with made-up alternatives using equivalence class substitution. *Birth* values were generalized to the year of birth. Three tuples have their *Ethnicity* values suppressed and 6 tuples have their *ZIP* code values valid to only the first 4 digits. For every assigned value of {*Ethnicity*, *Birth*, *Sex*, *ZIP*} found in the released table, there are at least 2 tuples having that value. The unique tuples (Caucasian, 10/23/64, Male, 02138) and (Caucasian, 3/15/65, female,02139) were modified. In the case of (Caucasian, 3/15/65, female,02139), the *Ethnicity* value was suppressed and *ZIP* generalized; and then, so that the suppressed value could not be inferred, a mutual suppression and generalization was performed on (Black, 8/24/65, female, 02138). A suppressed value must also adhere to the *k* requirement, so (Black, 10/23/65, female, 02138) was modified.

A comparison of the precision of the results from Datafly (in Figure 29), µ-Argus (in Figure 30), and k-Similar (in Figure 33) is provided in Figure 32. Datafly had less precision than µ-Argus, but

provided adequate protection. In comparison, k-Similar provided adequate protection also, but maintained more precision.

System	Precision	*k*-anonymity
Datafly	0.71	yes
µ-Argus	0.86	no
k-Similar	0.80	yes

Figure 34 Precision measurements for Datafly, µ-Argus and *k*-Similar

k-Similar as an anonymous data system

k-Similar uses the following disclosure limitation techniques: de-identification, equivalence class substitution, generalization, and suppression. Below is a description of the framework in which *k*-Similar operates.

S = {*subjects whose information is included in PT*}

P = *set of all people whose information could possibly be in PT*

PT = *privately held information about S*

QI = set of attributes with replications in E

U = P

RT = *k-Similar* (PT)

E = set of publicly available information in today's society

G = set of standard communication methods.

f = k-Similar

The system **A**(S, P, PT, *QI*, U, {RT}, E, G, *k-Similar*) is an $\mathbf{ADS_0}$.

Informal proof.

Let PT = data in Figure 28.

There cannot exist fewer than *k* tuples in RT having the same values across *QI*

based on the correctness of the *k*-Similar clustering algorithm.

So, *k*-map protection is provided and **A** is an $\mathbf{ADS_0}$.

Despite the fact that more specificity remains in the resulting data from *k*-Similar, making it more useful to the recipient, the underlying issues remain the same as concerning the correct identification of quasi-identifiers and the selection of *k*. Possible remedies to these problems are provided in the next section.

The precision metric introduced by *k*-Similar is useful in general for measuring the distortion of public-use files and can be extended to work with any combination of disclosure limitation techniques described earlier (not just generalization and suppression). In a preliminary survey of publicly available hospital discharge data, I found these data were 50% more distorted than necessary while still in many cases providing inadequate protection. This means that the data are not useful as possible so researchers have to make arrangements to get the more sensitive version. The overall impact is that more copies of sensitive data are distributed than necessary and at the same time, the pubic-use version of the data are typically still not sufficiently anonymous.

7 Discussion

The Scrub System demonstrated that medical data, including textual documents, can be de-identified, but as I have shown, de-identification alone is not sufficient to protect confidentiality. Not only can de-identified information often be re-identified by linking data to other databases, but also releasing too many patient-specific facts can identify individuals. Unless society is proactive, the proliferation of medical data may become so widespread that it will be impossible to release medical data without further breaching confidentiality. For example, the existence of rather extensive registers of business establishments in the hands of government agencies, trade associations and firms like Dunn and Bradstreet has virtually ruled out the possibility of releasing database information about businesses [41].

The Datafly, μ-Argus and *k*-Similar systems illustrated that medical information can be generalized so that attributes and combinations of attributes adhere to a minimal *k* requirement, and by so doing, confidentiality can be maintained. Such schemes can provide anonymous data for public use. There are drawbacks to these systems, but the primary shortcomings may be counteracted by policy.

One concern with both μ-Argus, Datafly and *k*-Similar is the determination of the proper value for *k* and its corresponding measure of disclosure risk. There is no standard that can be applied to assure that the final results are adequate. It is customary to measure risk against a specific compromising technique, such as linking to known databases that the data holder assumes the recipient is using. Several researchers have proposed mathematical measures of the risk, which compute the conditional probability of the linker's success [42].

A policy could be mandated that would require the producer of data released for public use to guarantee with a high degree of confidence that no individual within the data can be identified using demographic or semi-public information. Of course, guaranteeing anonymity in data requires a criterion against which to check resulting data and to locate sensitive values. If this is based only on the database itself, the minimum *k* and sampling fractions may be far from optimal and may not reflect the general population. Researchers have developed and tested several methods for estimating the percentage of unique values in the general population based on a smaller database [43]. These methods are based on subsampling techniques and equivalence class structure. In the absence of these techniques, uniqueness in the population based on demographic attributes can be determined using population registers that include patients from the database, such as local census data, voter registration lists, city directories, as well as information from motor vehicle agencies, tax assessors and real estate agencies. To produce an anonymous database, a producer could use population registers to identify sensitive demographic values within a database, and thereby obtain a measure of risk for the release of the data.

The second drawback with the μ-Argus, Datafly and *k*-Similar systems concerns the dichotomy between researcher needs and disclosure risk. If data are explicitly identifiable, the public expects patient permission to be required. If data are released for public use, then the producer must guarantee, with a high degree of confidence, that the identity of any individual cannot be determined using standard and predictable methods and reasonably available data. But when sensitive de-identified, but not necessarily anonymous, data are to be released, the likelihood that an effort will be made to re-identify an individual increases based on the needs of the recipient, so any such recipient has a trust relationship with society and the producer of the data. The recipient should therefore be held accountable.

The Datafly, *k*-Similar and μ-Argus systems quantify this trust by having the data holder identify quasi-identifiers among the attributes requested by the recipient. But recall that the determination of a quais-identifier requires guesswork in identifying attributes on which the recipient could link. Suppose a quasi-identifier is incorrect; that is, the producer misjudges which attributes are sensitive for linking. In this case, the Datafly, *k*-Similar and μ-Argus systems might release data that are less anonymous than what was required by the recipient, and as a result, individuals may be more easily identified. This risk cannot be perfectly resolved by the producer of the data since the producer cannot always know what resources the recipient holds. The obvious demographic attributes, physician identifiers, and billing information attributes can be consistently and reliably protected. However, there are too many sources of semi-public and private information such as pharmacy records, longitudinal studies, financial records, survey responses, occupational lists, and membership lists, to account a priori for all linking possibilities.

What is needed is a contractual arrangement between the recipient and the producer to make the trust explicit and share the risk. Figure 35 contains some guidelines that make it clear which attributes need to be protected against linking. Using this additional knowledge and the techniques presented in the Datafly, *k*-Similar and μ-Argus systems, the producer can best protect the anonymity of patients in data even when sensitive information is released. It is surprising that in most releases of medical data there are no contractual arrangements to limit further dissemination or use of the data. Even in cases where there is an IRB review, no contract usually results. Further, since the harm to individuals can be extreme and irreparable and can occur without the individual's knowledge, the penalties for abuses must be stringent. Significant sanctions or penalties for improper use or conduct should apply since remedy against abuse lies outside technology and statistical disclosure techniques and resides instead in contracts, laws and policies.

1. There must be a legitimate and important research or administrative purpose served by the release of the data. The recipient must identify and explain which attributes in the database are needed for this purpose.

2. The recipient must be strictly and legally accountable to the producer for the security of the data and must demonstrate adequate security protection.

3. The data must be de-identified. The release must contain no explicit individual identifiers nor should it contain data that would be easily associated with an individual.

4. Of the attributes the recipient requests, the recipient must identify which of these attributes, during the specified lifetime of the data, the recipient could link to other data the recipient will have access to, whether the recipient intends to link to such data or not. The recipient must also identify those attributes for which the recipient will link the data. If such linking identifies patients, then patient consent may be warranted.

5. The data provider should have the opportunity to review any publication of information from the data to insure that no potential disclosures are published.

6. At the conclusion of the project, and no later than some specified date, the recipient must destroy all copies of the data.

7. The recipient must not give, sell, loan, show or disseminate the data to any other parties.

Figure 35. Contractual requirements for restricted use of data based on federal guidelines and the Datafly System.

In closing this paper, a few researchers may not find this presentation of the magnitude and scope of the problem surprising, but it has disturbed legislators, scientists and federal agencies [44], so much so, I warn against overreaction especially as it may lead to inappropriate and inoperable policies. I present the problem and these incremental solutions from a belief that knowledge and not ignorance provides the best foundation for good policy. What is needed is a rational set of disclosure principles,

which are unlikely to evolve from piecewise reactions to random incidents, but require instead comprehensive analysis of the fundamental issues. The technology described here is quite helpful, but society must still make conscious decisions. There is a danger in over-simplifying this work. It does not advocate giving all the data on all the people without regard to whether individuals can be identified. It does not advocate releasing data that is so general it cannot be useful; substantial suppression does not appear to be the norm. From the viewpoint of the person who is to receive the data, these systems seek to provide the most general data possible that is practically useful. From the viewpoint of privacy, if that level of generality does not provide sufficient protection, then the techniques presented here identify the nature and extent of trust required for a given release of data. Polices and regulations regarding the agreements necessary to make that trust explicit and enforce its terms lie outside the technology.

Consider the case of data released to researchers. When anonymous data is useful, then the data should be released. In some cases completely anonymous data is not practically useful; in those cases, society (and the data holder) can quantify the trust given to researchers who receive more identifiable data. Changes should be made such that public-use files adhere to a reasonably high level of anonymity. In cases where more identifiable data is needed, society should consciously decide how to release such data and the recipient should be held responsible not to violate the contractual agreements that spell out the conditions of trust.

Finally I also warn against doing nothing. Consider an alternative to autonomous database systems, since the burden of determining the risk of disclosure may appear cumbersome. Suppose instead that society had a centralized federal repository for medical data like those found in Canada and other countries. Though institutions and businesses could maintain their own data for internal purposes, they could not sell or give data away in any form, except of course for disclosure to the federal repository, remuneration for services and required reporting. The recipients of these data would, in turn, be equally restricted against further dissemination. The trusted authority that maintains the central repository would have nearly perfect omniscience and could confidently release data for public use. Questions posed by researchers, administrators and others could be answered without releasing any data; instead the trusted authority would run desired queries against the data and then provide non-compromising results to the investigators.

In releases of de-identified data, the exact risk could be computed and accompanying penalties for abuse incorporated into the dissemination process. While this type of system may have advantages to maintaining confidentiality, it requires a single point of trust or failure. Current societal inclinations suggest that the American public would not trust a sole authority in such a role and would feel safer with distributed, locally controlled data. Ironically, if current trends continue, a handful of independent

information brokers may assume this role of the trusted authority anyway. If information brokers do emerge as the primary keepers of medical data (akin to the function that Dunn and Bradstreet serve for business data) they may eventually rank among the most conservative advocates for maintaining confidentiality and limiting dissemination. Their economic survival would hinge on protecting what would be their greatest asset, our medical records.

References

1 Kohane et al., "Sharing Electronic Medical Records Across Heterogeneous and Competing Institutions," in J. Cimino, ed., *Proceedings, American Medical Informatics Association* (Washington, D.C.: Hanley & Belfus, 1996):608-12.

2 Office of Technology Assessment, *Protecting Privacy in Computerized Medical Information* (Washington, D.C.: U.S. Government Printing Office, 1993).

3 See L.O. Gostin et al., "Privacy and Security of Personal Information in a New Health Care System," *Journal of the American Medical Association*, 270 (1993): at 2487 (citing Louis Harris and Associates, *The Equifax Report on Consumers in the Information Age* (Atlanta: Equifax, 1993)).

4 Louis Harris and Associates, *The Equifax-Harris Consumer Privacy Survey* (Atlanta: Equifax, 1994).

5 G. Cooper et al., "An evaluation of Machine-Learning Methods for Predicting Pneumonia Mortality," *Artificial Intelligence in Medicine*, 9, no. 2 (1997):107-38.

6 B. Woodward, "Patient Privacy in a Computerized World," *1997 Medical and Health Annual* (Chicago: Encyclopedia Britannica, 1996):256-59.

7 National Association of Health Data Organizations, *A Guide to State-Level Ambulatory Care Data Collection Activities* (Falls Church: National Association of Health Data Organizations, Oct. 1996).

8 P. Clayton et al., National Research Council, *For the Record: Protecting Electronic Health Information* (Washington, D.C.: National Academy Press, 1997).

9 See, for example, Donna E. Shalala, Address at the National Press Club, Washington, D.C. (July 31, 1997).

10 D. Linowes and R. Spencer, "Privacy: The Workplace Issue of the '90s," *John Marshall Law Review*, 23 (1990):591-620.

11 D. Grady, "Hospital Files as Open Book," *New York Times*, March 12, 1997, at C8.

12 "Who's Reading Your Medical Records," *Consumer Reports*, October (1994): 628-32.

13 See note 7 National Association of Health Data Organizations.

14 Group Insurance Commission testimony before the Massachusetts Health Care Committee. See *Session of the Joint Committee on Health Care, Massachusetts State Legislature*, (March 19, 1997).

15 Cambridge Voters List Database. *City of Cambridge, Massachusetts*. Cambridge: February 1997.

16 See note 14 Group Insurance Commission.

17 J. Ullman. *Principles of Database and Knowledge Base Systems*. Computer Science Press, Rockville, MD. 1988.

18 I. Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, 1972, pp. 7-18.

19 T. Su and G. Ozsoyoglu. Controlling FD and MVD inference in multilevel relational database systems. *IEEE Transactions on Knowledge and Data Engineering*, 3:474--485, 1991.

20 M. Morgenstern. Security and Inference in multilevel database and knowledge based systems. *Proc. of the ACM SIGMOD* Conference, pages 357--373, 1987.

21 T. Hinke. Inference aggregation detection in database management systems. In *Proc. of the IEEE Symposium on Research in Security and Privacy*, pages 96-107, Oakland, 1988.

22 T. Lunt. Aggregation and inference: Facts and fallacies. In *Proc. of the IEEE Symposium on Security and Privacy*, pages 102--109, Oakland, CA, May 1989.

23 X. Qian, M. Stickel, P. Karp, T. Lunt, and T. Garvey. Detection and elimination of inference channels in multilevel relational database systems. In *Proc. of the IEEE Symposium on Research in Security and Privacy*, pages 196--205, 1993.

24 T. Garvey, T. Lunt and M. Stickel. Abductive and approximate reasoning models for characterizing inference channels. *IEEE Computer Security Foundations Workshop*, **4**, 1991.

25 D. Denning and T. Lunt. A multilevel relational data model. In *Proc. of the IEEE Symposium on Research in Security and Privacy*, pages 220-234, Oakland, 1987.

26 See note 21 Hinke.

27 See note 22 Lunt.

28 See note 23 Qian, Stickel, Karp, Lunt and Garvey.

29 D. Denning. *Cryptography and Data Security*. Addison-Wesley, 1982.

30 D. Denning, P. Denning, and M. Schwartz. The tracker: A threat to statistical database security. *ACM Trans. on Database Systems*, 4(1):76--96, March 1979.

31 G. Duncan and S. Mukherjee. Microdata disclosure limitation in statistical databases: query size and random sample query control. In *Proc. of the 1991 IEEE Symposium on Research in Security and Privacy*, May 20-22, Oakland, California. 1991.

32 T. Dalenius. Finding a needle in a haystack – or identifying anonymous census record. *Journal of Official Statistics*, 2(3):329-336, 1986.

33 G. Smith. Modeling security-relevant data semantics. In *Proceedings of the 1990 IEEE Symposium on Research in Security and Privacy*, May 1990.

34 I. Kohane, "Getting the Data In: Three-Year Experience with a Pediatric Electronic Medical Record System," in J. Ozbolt, ed., *Proceedings, Symposium on Computer Applications in Medical Care* (Washington, D.C.: Hanley & Belfus, 1994): 457-61.

35 G. Barnett, "The Application of Computer-Based Medical-Record Systems in Ambulatory Practice," *N. Engl. J. Med.*, 310 (1984): 1643-50.

36 Anon., Privacy & Confidentiality: Is It a Privilege of the Past?, Remarks at the Massachusetts Medical Society's Annual Meeting, Boston, Mass. (May 18, 1997).

37 Government Accounting Office, *Fraud and Abuse in Medicare and Medicaid: Stronger Enforcement and Better Management Could Save Billions* (Washington, D.C.: Government Accounting Office, HRD-96-320, June 27, 1996).

38 See note 14 Group Insurance Commission.

39 A. Hundepool and L. Willenborg, "μ- and τ-Argus: Software for Statistical Disclosure Control," *Third International Seminar on Statistical Confidentiality* (1996) (available at http://www.cbs.nl/sdc/argus1.html.

40 For a presentation of the concepts on which μ-Argus is based, see L. Willenborg and T. De Waal, *Statistical Disclosure Control in Practice* (New York: Springer-Verlag, 1996).

41 N. Kirkendall et al., *Report on Statistical Disclosure Limitation Methodology*, *Statistical Policy Working Paper* (Washington, D.C.: Office of Management and Budget, no. 22, 1994).

42 G. Duncan and D. Lambert, "The Risk of Disclosure for Microdata," *Proceedings of the Bureau of the Census Third Annual Research Conference* (Washington, D.C.: Bureau of the Census, 1987): 263-74.

43 C. Skinner and D. Holmes, "Modeling Population Uniqueness," *Proceedings of the International Seminar on Statistical Confidentiality* (Dublin: International Statistical Institute, 1992): 175-99.

44 For example Latanya Sweeney's testimony before the Massachusetts Health Care Committee had a chilling effect on the proceedings that postulated that the release of deidentified medical records provided anonymity. See *Session of the Joint Committee on Health Care, Massachusetts State Legislature*, (Mar. 19, 1997) (testimony of Latanya Sweeney, computer scientist, Massachusetts Institute of Technology). Though the Bureau of the Census has always been concerned with the anonymity of public use files, they began new experiments to measure uniqueness in the population as it relates to public use files. Computer scientists who specialize in database security are re-examining access models in light of these works.