Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Releasing Individually Identifiable Microdata with Privacy Protection Against Stochastic Threat: An Application to Health Information

Robert Garfinkel, Ram Gopal, Steven Thompson,

To cite this article:
Robert Garfinkel, Ram Gopal, Steven Thompson, (2007) Releasing Individually Identifiable Microdata with Privacy Protection Against Stochastic Threat: An Application to Health Information. Information Systems Research 18(1):23-41. https://doi.org/10.1287/isre.1070.0112

Information Systems Research
Vol. 18, No. 1, March 2007, pp. 23–41
ISSN 1047-7047 | EISSN 1526-5536 | 07 | 1801 | 0023

informs

DOI 10.1287/isre.1070.0112

Releasing Individually Identifiable Microdata with Privacy Protection Against Stochastic Threat: An Application to Health Information

Robert Garfinkel, Ram Gopal, Steven Thompson
Department of Operations and Information Management, School of Business,
University of Connecticut, Storrs, Connecticut 06029
{rob.garfinkel@business.uconn.edu, ram.gopal@business.uconn.edu, sthomps3@richmond.edu}

The ability to collect and disseminate individually identifiable microdata is becoming increasingly important in a number of arenas. This is especially true in health care and national security, where this data is considered vital for a number of public health and safety initiatives. In some cases legislation has been used to establish some standards for limiting the collection of and access to such data. However, all such legislative efforts contain many provisions that allow for access to individually identifiable microdata without the consent of the data subject. Furthermore, although legislation is useful in that penalties are levied for violating the law, these penalties occur after an individual's privacy has been compromised. Such deterrent measures can only serve as disincentives and offer no true protection. This paper considers security issues involved in releasing microdata, including individual identifiers. The threats to the confidentiality of the data subjects come from the users possessing statistical information that relates the revealed microdata to suppressed confidential information. The general strategy is to recode the initial data, in which some subjects are "safe" and some are at risk, into a data set in which no subjects are at risk. We develop a technique that enables the release of individually identifiable microdata in a manner that maximizes the utility of the released data while providing preventive protection of confidential data. Extensive computational results show that the proposed method is practical and viable and that useful data can be released even when the level of risk in the data is high.

Key words: data security; privacy; health information; optimization
History: Sumit Sarkar, Senior Editor; Ramayya Krishnan, Associate Editor. This paper was received on February 28, 2005, and was with the authors 6 months for 2 revisions.

1. Introduction

As information storage and processing capabilities increase, a number of groups and organizations are engaging in the collection and dissemination of individually identifiable microdata (IIM). Examples include the Department of Homeland Security, the Centers for Disease Control and Prevention, insurance companies, and various state and local public health departments. In some cases IIM are collected and used by a specific organization. In other cases data is collected and shared with other organizations. The collection and dissemination of IIM is typically considered justifiable when the objectives of the data recipient are deemed to be "for the greater good" and statistical data alone is not sufficient to achieve those objectives.

In recognition of the fact that IIM is highly sensitive, especially in relation to matters such as medical or financial information, a number of laws have been passed that address the question of when IIM can be collected and shared. Examples at the federal level include the Privacy Act of 1974, the Computer Matching and Privacy Protection Act of 1988, the Paperwork Reduction Act of 1995, the Principles for Providing and Using Personal Information ("Privacy Principles"), published by the Information Infrastructure Task Force in 1995, and the Health Insurance Portability and Accountability Act (HIPAA), enacted in 1996. In most cases these laws provide substantial disincentives for the abuse of IIM. For instance, the maximum penalty under HIPAA for the abuse of personal health information is a $250,000 fine and up to 10 years imprisonment.

Nevertheless, although enacted for the purpose of protecting individual privacy in the face of an increasingly computerized world, all these laws contain

provisions that allow for the collection and dissemination of IIM. The HIPAA Privacy Rule provides a good example of such provisions as related to medical information. The following summary, taken from the CDC website (2005) describes the current situation well:

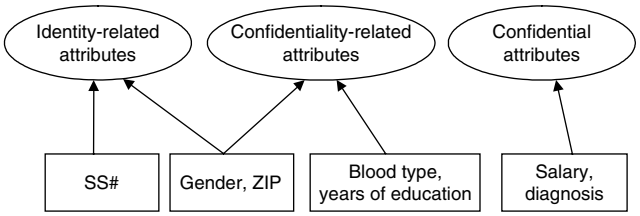
> New national health information privacy standards have been issued by the U.S. Department of Health and Human Services (DHHS), pursuant to the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The new regulations provide protection for the privacy of certain individually identifiable health data, referred to as protected health information (PHI). Balancing the protection of individual health information with the need to protect public health, the Privacy Rule expressly permits disclosures without individual authorization to public health authorities authorized by law to collect or receive the information for the purpose of preventing or controlling disease, injury, or disability, including but not limited to public health surveillance, investigation, and intervention.

Legal protective measures that contain these types of provisions can be problematic for at least two reasons. First, there is often ambiguity regarding what qualifies a given situation as being "exempt" from the law. The result is that in some cases IIM is "unjustifiably" collected and disseminated and in others a given initiative might be incorrectly cancelled under the mistaken belief that the collection and dissemination of IIM is not justified.

The second concern is that legal protection ultimately takes the form of a deterrent measure and the data—and therefore the data subjects—are essentially unprotected. That is, under existing law, if IIM is abused and *if* this abuse is detected and *if* the abuse can be ascribed to certain individuals in a court of law, then the perpetrators will be punished. This may come as small consolation to data subjects who may have suffered loss of employment or employment opportunities, cancelled insurance policies, or public embarrassment as a result of the unlawful use of IIM.

Confidentiality considerations typically center around three categories of attributes, as described in Figure 1. Confidential attributes (e.g., salary or medical diagnosis) are sensitive information of the respondents or subjects in the data set. A central responsibility of the data provider is to ensure that

Figure 1 Categorization of Data Set Attributes



the confidential information on any of the subjects is not divulged to the data users. Identity-related attributes are those that can directly or indirectly help a user identify an individual or subject in the data set. These may be such powerful indicators as name or social security number, or much weaker ones such as profession or city of domicile. Confidentiality-related attributes are not confidential in and of themselves but are statistically related to the confidential attributes. Information on these attributes may allow a user to estimate confidential values even if they have been suppressed.

The categorization of identity-related and confidentiality-related attributes may not be mutually exclusive. For instance, identity-related attributes such as gender and age may also exhibit statistically significant relationships with a confidential attribute such as a particular medical condition.

It is common to release "de-identified" data in formats that are useful for statistical users. This can be achieved in any number of ways, as indicated in the literature review of the next section. Here we consider how to release data in a safe and useful manner for "IIM users." By their nature these released data will be consistent with any data released for statistical purposes. Also there is no need to make any distinction between a "data snooper" and a "legitimate" user. We assume that the subjects of the database desire to have their confidential information protected against any and all users.

When considering the release of IIM that contain a confidential attribute, one obvious approach to protecting the confidentiality of the data subjects would be to simply remove the confidential attribute from the released data. This is an important step, but it is not sufficient because it is possible to infer the confidential field value based on statistical data that relate the released confidentiality-related field values to the confidential field value.

The objective of this research is to develop a method that allows the release of IIM while minimizing information loss and, at the same time, providing a degree of preventive confidentiality protection to the data subjects. It is important to note that this technique is not meant to be a replacement for existing legal measures. Rather, it is a supplement designed to offer a degree of preventive protection for confidential information that is not currently available.

The remainder of the paper is organized as follows. Section 2 contains a review of related literature. Section 3 provides a conceptual overview of the model and solution technique and introduces a motivating example. Decision variables and various stochastic measures are introduced in §4, and §5 is devoted to precise definitions of risk. Section 6 contains the linear programming model, and §7 discusses the extensive computational results, as well as implementation issues and managerial implications. Conclusions and future research are presented in §8.

2. Related Literature

The information systems literature contains a considerable amount of research in the general vein of information security and confidentiality protection. A great deal of research effort has been dedicated to the problem of maximizing the amount and the utility of data that can be released without jeopardizing the individual's right to privacy (see, e.g., Adam and Wortman 1989, Chowdhury et al. 1999, Gopal et al. 1998, Garfinkel et al. 2002, Muralidhar et al. 1995). A number of techniques have been developed to address the problem of maximizing information provision while ensuring that the revealed information does not permit a user to infer confidential data about any individual subject. Suppression of cells in the tables to be released is a common strategy (Cox 1980, Carvalho et al. 1994, Causey et al. 1985, Duncan et al. 2003a, Fischetti and Salazar 2001, Kelly 1990, Geurts 1992, Zayatz 1992, Cox 1992). Other methods include row and column aggregation (Willenborg and Hundepool 1998), data perturbation (Muralidhar et al. 1995, Duncan and Fienberg 1999, Sarathy and Muralidhar 2002), and camouflage (Gopal et al. 1998, Garfinkel et al. 2002).

A number of software products, including μ and τ-Argus (Hundepool and Willenborg 1996) and Datafly (Sweeney 2002a), have also been developed to enable safe dissemination of data products. The increased use of data-mining tools also presents a set of unique confidentiality problems that the academic literature has begun to address (Li and Sarkar 2006a, b; Menon et al. 2005).

In all this work, the threat to confidentiality generally comes down to the identity disclosure problem, i.e., whether the user can positively link the identity of an individual with a set of variable values (Dobra et al. 2003). A standard precaution is to sanitize the data by eliminating attributes that directly lead to the identity of the data subjects (e.g., name or social security number). The resulting data set is presented in a format that enables the users to obtain statistical information on the remaining attributes, but prevents a user from associating confidential data with subjects. Recent research has explored the risk of identity disclosure in the released data set that may persist even after the elimination of attributes that directly lead to the identity of data subjects. The fundamental issue addressed in the k-anonymity problem is the generation of data sets in which each subject is indistinguishable from no fewer than k other subjects (Domingo-Ferrer and Mateo-Sanz 2002, Sweeney 2002b). The k-anonymity problem has been shown to be NP-hard (Meyerson and Williams 2004), and fast techniques that yield good, but not necessarily optimal, solutions have been the subject of a great deal of research (Aggarwal 2005, Aggarwal et al. 2005, Domingo-Ferrer and Tora 2005). Work on k-anonymity has also been applied to more specific problem settings such as locational privacy (Gedik and Liu 2005), distributed databases (Jiang and Clifton 2005), and facial derecognition (Newton et al. 2005).

Regardless of the application area, the fundamental problem is the same, preventing the reidentification of a subject to within k records. The techniques developed in this paper differ from prior work in that the released data set explicitly includes the identity of the data subjects. We extend the general philosophy of prior research, that of transforming data prior to their release to protect the data subjects, to the domain of IIM. In this setting the "R-U confidentiality map" (Duncan et al. 2003b) is adopted as a framework for capturing the fundamental trade-off between mitigating risk and the utility of the released data.

Table 1 Individually Identifiable Microdata

Name	Postal code	Gender	C-Reactive protein	Cholesterol	Blood pressure	Glucose	Diagnosis
M. A.	06040	M	H	H	H	V	{Diabetes, Heart, MRSA}
G. P.	06269	M	H	H	H	V	{Diabetes, Heart}
M. L.	14260	F	H	H	H	V	{Diabetes, Heart}
W. F.	14260	M	L	H	L	N	
R. H.	06040	F	L	H	L	N	
F. J.	06269	M	N	N	N	N	
M. G.	98195	F	N	N	N	N	Heart
J. M.	98195	F	N	N	N	N	
A. B.	98195	F	N	N	N	N	
J. R.	14260	M	N	N	N	H	Diabetes
R. S.	98195	M	N	N	N	H	{Diabetes, Heart}
R. G.	90210	M	L	L	L	N	
S. T.	23059	M	L	L	L	N	
J. T.	23059	F	N	N	V	N	Heart
M. D.	44187	F	N	N	V	N	

3. Recoding Individually Identifiable Microdata: Conceptual Overview

The technique described in this paper can be thought of as a special type of recoding (e.g., μ-Argus, see Hundepool and Willenborg 1996). In conventional recoding the values of individual attributes are merged attribute by attribute. The method presented in this paper involves first grouping attribute values into sets, and then merging over these sets. To illustrate, consider the problem faced by a data provider in possession of the individually identifiable medical data set depicted in Table 1. Table 1 contains data according to the following abbreviations: gender {Male (M), Female (F)}, C-reactive protein (CRP) levels {Low (L), Normal (N), High (H)}, cholesterol {Low (L), Normal (N), High (H)}, blood pressure (BP) {Low (L), Normal (N), High (H), Very High (V)}, and serum glucose (Gluc) levels {Normal (N), High (H), Very High (V)}. The field "Diagnosis" is considered to be the confidential field. The data depicted in Table 1 will be used as an illustrative example throughout the paper. Additional, more detailed application areas in public health and data mining are provided in the e-companion.[1]

To ensure protection of confidential information on the subjects represented in the data set, grouped and de-identified data is normally released to satisfy the information needs of statistical users. This is illustrated in Table 2, where all individual identifiers are removed and aggregate information on the confidential attributes, grouped by confidentiality-related attribute values, is released. Such a table, if deemed safe by the data provider, will be released in our model along with the perturbed database to be developed in the remainder of the paper.

We define an *input channel* as a combination of confidentiality-related attribute values for which confidential information is released. For example, C-Reactive protein = H, cholesterol = H, blood pressure = H, glucose = V, would be represented as the input channel {H, H, H, V}.

The information in Table 2 enables users to identify trends and high-risk groups. For example, from Table 2 it is possible to determine that those with high levels of C-Reactive protein, cholesterol, and blood pressure as well as very high serum glucose levels, i.e., the subjects described by input channel {H, H, H, V}, are at greater risk for heart disease, diabetes, and a dangerous infection called methicillin-resistant Staphylococcus aureus (MRSA) than the general population (which in this illustration is quite small for the purpose of having a viable working example).

The ability to determine that those particular subjects are at high risk for those diseases is important. However, the statistical information alone is not enough for public health workers who may also want to contact those subjects. For instance, a public health worker may want to notify those at risk for diabetes and heart disease of new dietary guidelines and information regarding how to identify and seek treatment for so-called "silent heart attacks" (a heart attack that

Table 2 Statistical Data

Confidentiality related	# Subjects	# Diabetes (%)	# Heart (%)	# {Diabetes, Heart} (%)	# {Diabetes, Heart, MRSA} (%)
H, H, H, V	3	0 (0)	0 (0)	2 (67)	1 (33)
L, H, L, N	2	0 (0)	0 (0)	0 (0)	0 (0)
N, N, N, N	4	0 (0)	1 (25)	0 (0)	0 (0)
N, N, N, H	2	1 (50)	0 (0)	1 (50)	0 (0)
L, L, L, N	2	0 (0)	0 (0)	0 (0)	0 (0)
N, N, V, N	2	0 (0)	1 (50)	0 (0)	0 (0)

[1] The e-companion to this paper is available on the *Information Systems Research* website at http://isr.pubs.informs.org/ecompanion.html.

is not accompanied by symptoms such as chest pain and difficulty breathing), which afflict diabetics more commonly than the general population. In this case the public health worker would want those at risk to understand the risk factors for silent heart attacks and advise them to be screened by a physician for silent ischemia, which is a precursor to a silent heart attack. Likewise, the ability to identify and contact those at risk for MRSA represents an important public health initiative.

In this work we show that this can be achieved while protecting the confidential information of the subjects. The information utility measure, corresponding to that of Duncan et al. (2003b), is inversely proportional to the number of spurious subjects who could potentially be contacted and told that they may be at risk. Such subjects should easily be able to verify that the warning does not pertain to them.

The application of the technique proposed here results in a modified version of the original microdata set with the individual identifiers intact and the confidential attributes removed. Further, it is based on *output channels*, which are defined as sets of one or more input channels. It contains as much information as is safely possible to reveal on the confidentiality-related fields. Table 3 illustrates, in essence, the format of the released IIM.

3.1. Mapping from Input Channels to Output Channels

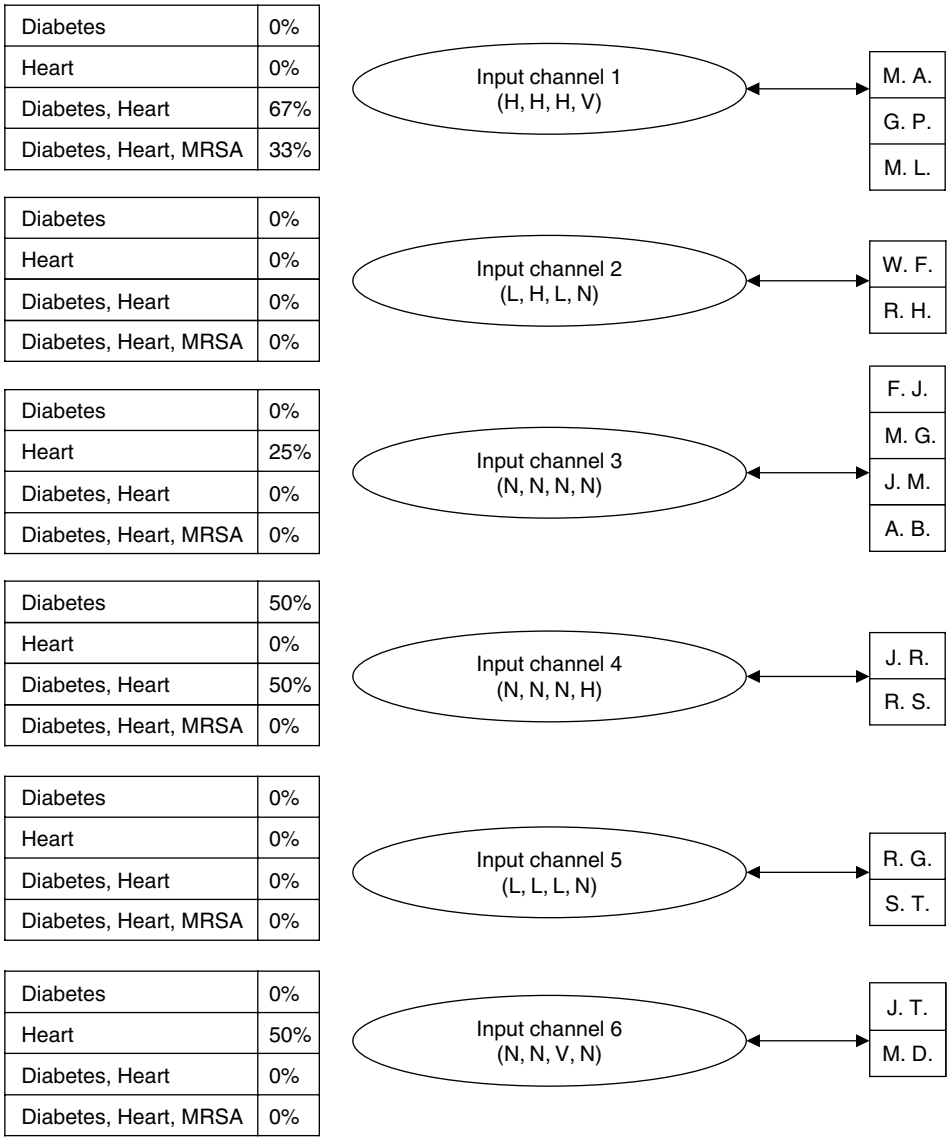
Figure 2 provides an input channel-based representation of the information in Table 1. Releasing the information as shown in Figure 2 is equivalent to "full revelation," in the sense that the data recipient knows exactly which input channel describes each subject.

Even though confidential information is not included explicitly here, it would be simple for a user to stochastically infer confidential information about the subjects described by a given input channel. This is illustrated in Figure 2, which shows, for each input channel, the percentage of subjects that have a particular disease. An extreme solution would be to protect the subjects by assigning all of them to a single output channel that contained, as its elements, all six input channels. Figure 3 illustrates the result.

Table 3 Individually Identifiable Microdata Format

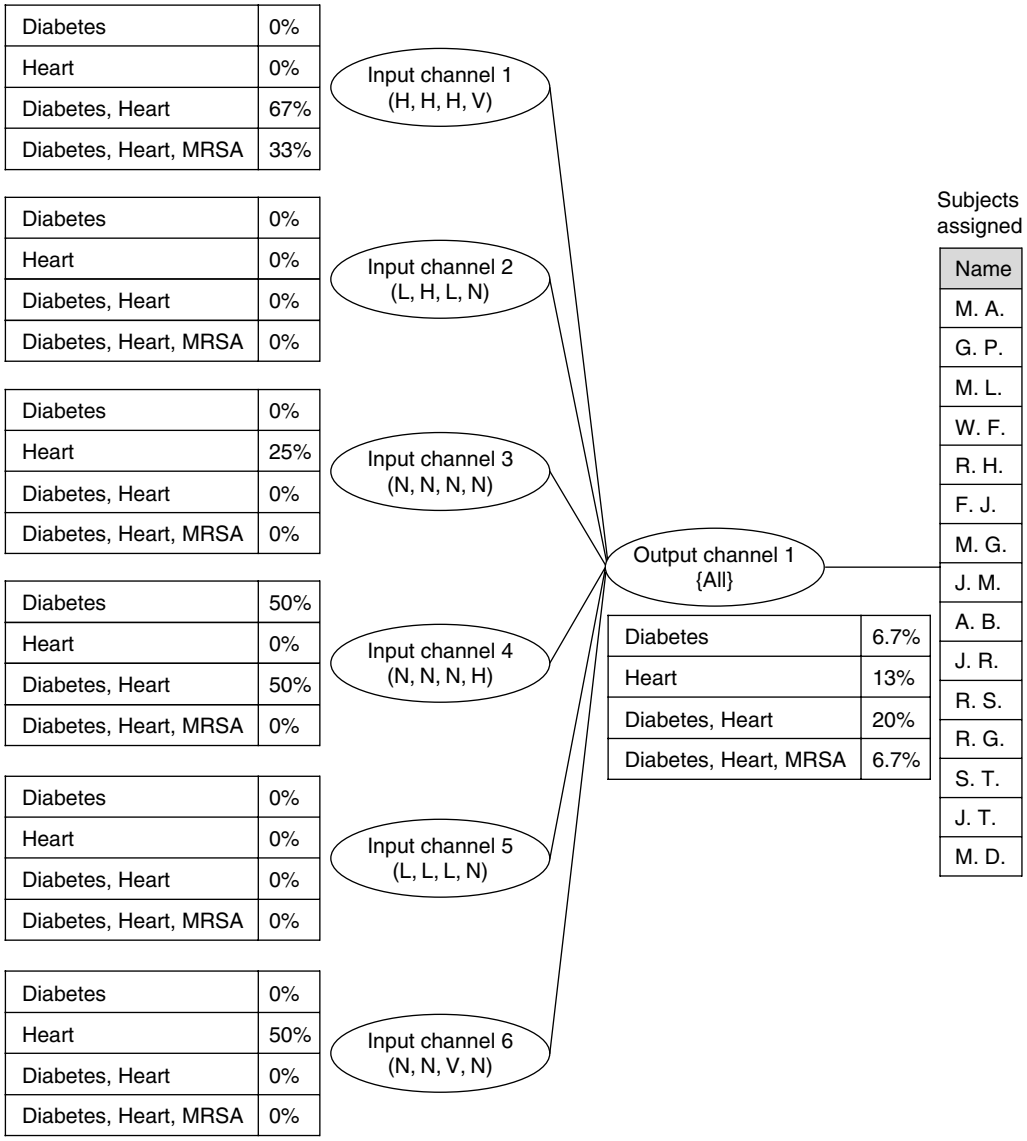
Name	ZIP	Gender	C-Reactive protein	Cholesterol	Blood pressure	Glucose	Diagnosis
M. A.	06040	M	As much as possible				Removed
G. P.	06269	M					
M. L.	14260	F					
W. F.	14260	M					
R. H.	06040	F					
F. J.	06269	M					
M. G.	98195	F					
J. M.	98195	F					
A. B.	98195	F					
J. R.	14260	M					
R. S.	98195	M					
R. G.	90210	M					
S. T.	23059	M					
J. T.	23059	F					
M. D.	44187	F					

Figure 2 Full Revelation



By assigning all subjects to an output channel comprised of all input channels, the data recipient is only able to determine that each subject described by that output channel is at the same risk for a given disease as the general population. Figure 3 illustrates a key condition that must be satisfied in order to use this technique. Clearly a necessary and sufficient condition for feasibility of the overall problem is that the output channel consisting of all input channels is "safe" (to be formally defined in §5). That is, mere membership in the original data set is not risky for any individual. Obviously, if this criterion is not satisfied, no microdata can be revealed. One solution to the overall problem of safely assigning subjects to output channels while minimizing total information loss is illustrated in Figure 3. However, that solution, even if safe, contains no useful information beyond that of Figure 2, except for the individual identifiers (IIDs). Since all subjects are assigned to a single output channel, they become indistinguishable and any query will result in a list that includes every subject in the data set.

Figure 3 Minimal Revelation



A third approach involves partial revelation. This is illustrated in Figure 4, where the data recipient still receives some information regarding which input channels do not describe a given subject (e.g., M. A. is not in input channels 4, 5, or 6).

Figure 4 reveals some general concepts that are important in the remainder of this work. First, it is always the case that the output channel to which a subject has been assigned has, as an element, the input channel that originally described the subject. This ensures that the output database is "inclusive," in the sense that any query on the released microdata for subjects meeting certain criteria will be guaranteed to yield a set of subjects that includes the set of subjects that would have resulted from the same query run on the original data set. Second, the data recipient is told the composition of each output channel, i.e., the input channels that comprise each output channel, is made known. Note that transformation from input to output channels can be considered simply expansion of the confidentiality-related fields. Furthermore, by performing this expansion at the "channel level" instead of the "attribute level," the user receives more useful information. Consider output channel 1 in

Figure 4 Partial Revelation

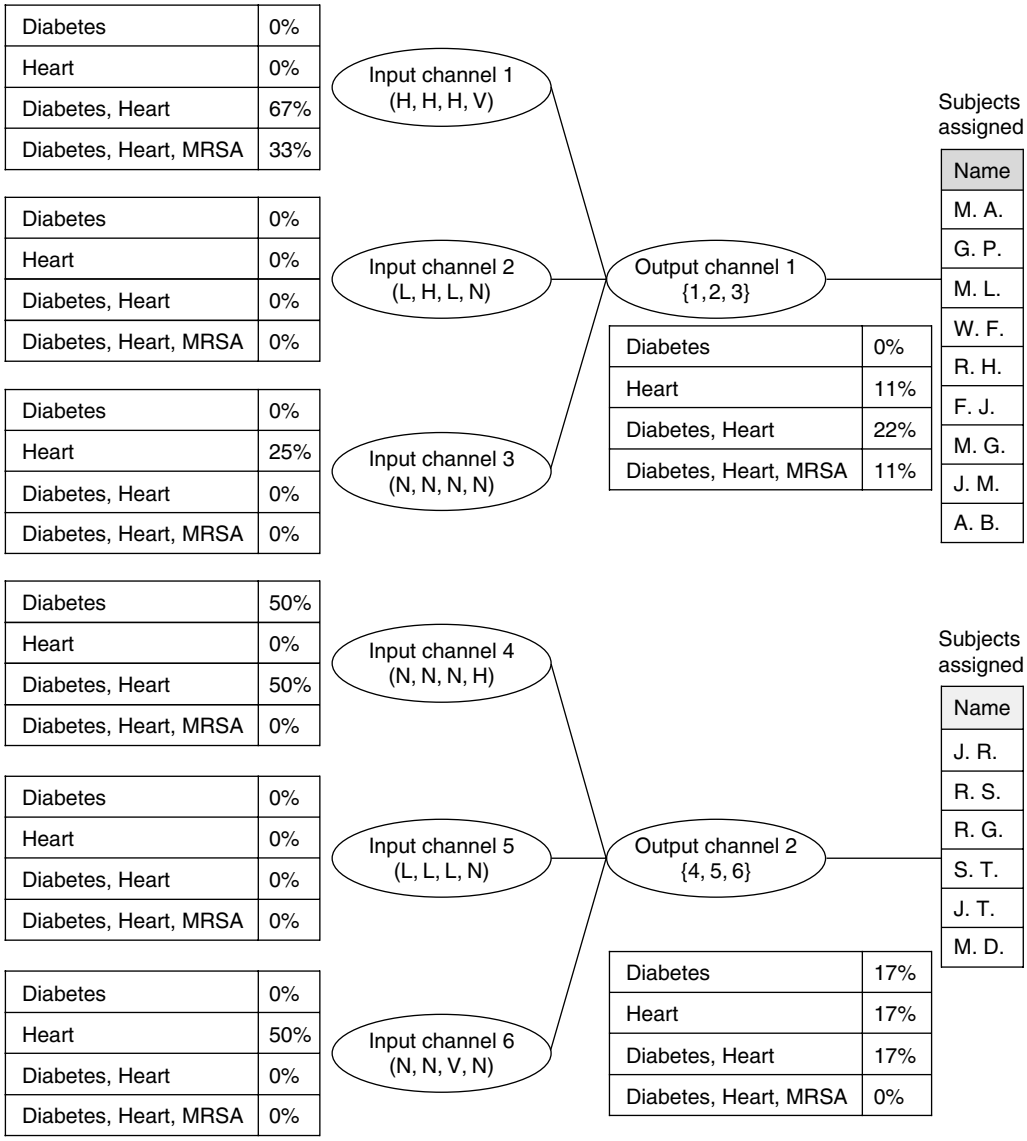


Figure 4. At the attribute level it would be ({L, N, H}, {N, H}, {L, N, H}, {N, V}), which could be considered to be the union of $3 \times 2 \times 3 \times 2 = 36$ input channels, as opposed to the three input channels of Figure 4. Thus, the data recipient receives much more precise information regarding which input channels might accurately describe the subjects and which do not.

The solution in Figure 4 results in a data set that, ignoring safety, is more useful than the one depicted in Figure 3 but less useful than that of Figure 2. To protect the confidentiality of the data subjects, some data utility was lost. Notice that the mapping, while "disguising" which input channel truly describes a subject, also results in the addition of spurious subjects to any query requesting subjects from a particular group. To illustrate how utility is measured in this context, consider a public health worker who wishes to contact subjects who are at risk for MRSA. First, assume that the data provider has decided that he or she does not want any user to be able to infer that a given "risky" value of the confidential field applies to any subject with "too great" a probability and that the mapping

depicted in Figure 4 is feasible (i.e., safe by that criterion). For instance, the data provider may not want any user to be able to infer that the confidential value "Diabetes and Heart" applies to any subject with a probability greater than 0.3.

Then suppose a public health worker determined from the statistical data in Table 2 that subjects described by input channel 1, i.e., {H, H, H, V}, represent the at-risk population in which he or she is interested. By using the perturbed individually identifiable microdata represented by Figure 4, a query for subjects described by input channel 1 would return a list of the nine subjects described by output channel 1. Three of these subjects truly meet that search criterion; the other six do not. Also, because the probability that a given subject in output channel 1 has "Diabetes and Heart" is 0.22, membership in output channel 1 would not pose a risk based on the bound of 0.3 established by the data provider. The public health worker also receives a guarantee that every subject he or she is interested in is included in the query output. The addition of spurious subjects results in data that has less utility to the users, and because the optimization model of §6 maximizes data utility, the number of spurious subjects added to queries will, in general, be as small as possible.

Clearly, the amount of risk that the data provider tolerates has a direct bearing on the overall utility of the released microdata set. This fundamental trade-off between risk and utility, typically framed in terms of an R-U confidentiality map (Duncan et al. 2003b), is an important decision aid for data providers to effectively create the IIM release strategy. The computational analysis reported in §7 provides important insights on the inherent nonlinearities in the risk-utility relationship.

4. The Channel Expansion Technique

We formally define an input channel $\mathbf{d}_r^{\text{in}}$ as the vector of confidentiality-related attribute values that describes the rth group of subjects. Let $\mathbf{D}^{\text{in}} := \{\mathbf{d}_r^{\text{in}}\}$ and $\mathbf{D}^{\text{out}} := \{\mathbf{d}_w^{\text{out}}\}$ be the set of input and output channels, respectively. Also let $V = \{v^k\colon k = 1, \ldots, K\}$ be the set of possible "risky" values, $v^0 =$ "none," and $V' = V \cup \{v^0\}$. We will often abuse notation by writing $k \in V$ or $k \in V'$ to mean $v^k \in V$ or $v^k \in V'$. The following notation and definitions will also be used:

- m: the number of individuals in the data set;
- m_r^{in}: the number of individuals in the input channel ds$\mathbf{d}_r^{\text{in}}$;
- $m_r^{\text{in},k}$: the number of individuals in $\mathbf{d}_r^{\text{in}}$ with confidential value v^k;
- $P(v^k)$: the probability that an individual in the data set has confidential value v^k. Clearly, $P(v^k) = (1/m)\sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{D}^{\text{in}}} m_r^{in,k}$;
- $P(v^k \mid \mathbf{d}_r^{\text{in}})$: the probability that an individual in $\mathbf{d}_r^{\text{in}}$ has confidential value v^k. It follows that $P(v^k \mid \mathbf{d}_r^{\text{in}}) = m_r^{in,k}/m_r^{\text{in}}$;
- m_w^{out}: the number of individuals in the output channel $\mathbf{d}_w^{\text{out}}$;
- $P(v^k \mid \mathbf{d}_w^{\text{out}})$: the probability that an individual in output channel $\mathbf{d}_w^{\text{out}}$ has confidential value v^k (see (1) below);
- x_{rw}: the number of individuals in input channel $\mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}}$ that are assigned to output channel $\mathbf{d}_w^{\text{out}}$.

The decisions to be made by the data provider are the composition of the output channels and how many subjects to assign from each input channel to each output channel. There are two advantages to making subject assignment decisions at the input channel level. That is, once the x_{rw} are determined the subjects are then randomly assigned from input channels to output channels, independent of their confidential field attribute values. First, the data provider will not be required to take into account the confidential field values of individual subjects. Second, by all subjects in a given input channel being treated as identical in terms of the distribution of the confidential field, the probability that an individual in output channel $\mathbf{d}_w^{\text{out}}$ has confidential value v^k is determined by

$$P(v^k \mid \mathbf{d}_w^{\text{out}}) = \frac{\sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}}} P(v^k \mid \mathbf{d}_r^{\text{in}}) x_{rw}}{\sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}}} x_{rw}}. \tag{1}$$

Finally, because the x_{rw} variables will, in general, be very large for realistic sized databases, there is no practical reason to impose integrality on them.

5. A General Model of Risk

5.1. Risky Input Channels

The first step is to partition the initial set of input channels $\mathbf{D}^{\text{in}}$ into two sets, risky and safe ($\mathbf{R}^{\text{in}}$, $\mathbf{S}^{\text{in}}$). An input channel can be deemed risky based on either

a single confidential field or a combination of confidential fields. In the first case, define $\mathbf{d}_r^{\text{in}} \in \mathbf{R}^{\text{in}}$ if $\mathbf{d}_r^{\text{in}}$ is k-risky, denoted $\mathbf{d}_r^{\text{in}} \in \mathbf{R}_k^{\text{in}}$, for some $k \in V$. Then $\mathbf{d}_r^{\text{in}} \in \mathbf{R}_k^{\text{in}}$ only if

$$P(v^k \mid \mathbf{d}_r^{\text{in}}) > P(v^k). \tag{2}$$

That is, an individual in $\mathbf{d}_r^{\text{in}}$ is considered potentially at risk if the likelihood that his value in the confidential field is higher than that of a subject chosen at random from the original data set. It is possible that even if (2) holds, the data provider may wish to set a higher threshold to define whether the potential threat is real enough to require action. An advantage of defining a higher threshold is that the released data set will ultimately have greater utility to the data recipient, because less perturbation will be required to make the data safe. The data provider can determine, value by value, the magnitude of the threat that must be present to cause concern. To do this, a constant $u_k \geq P(v^k)$ is established by the data provider for every v^k, such that $\mathbf{d}_r^{\text{in}} \in \mathbf{R}_k^{\text{in}}$ if and only if

$$P(v^k \mid \mathbf{d}_r^{\text{in}}) > u_k \geq P(v^k), \quad k \in V. \tag{3}$$

In the second case the data provider may be concerned, not only with whether the subjects in an input channel are at risk for a given value of the confidential field, but also with some measure of "how much" risk, aggregated over the confidential field values, can be assigned to subjects described by that channel. The data provider may feel that, even though (3) has not been satisfied for any k, the subjects in that channel may still be at risk. In fact, decisions involving such factors as employment or insurance coverage can be influenced by such "aggregate risk." Thus the data provider can assign a "weight" c_k to every $k \in V'$, which indicates how much of a relative threat that element is considered, where c_0 is set to zero. Then the second criterion for $\mathbf{d}_r^{\text{in}} \in \mathbf{R}^{\text{in}}$ is that the total weighted risk in $\mathbf{d}_r^{\text{in}}$ exceeds some parameter b established by the data provider. That is, $\mathbf{d}_r^{\text{in}} \in \mathbf{R}^{\text{in}}$ if

$$\sum_{k \in V'} P(v^k \mid \mathbf{d}_r^{\text{in}}) c_k > b. \tag{4}$$

Then $\mathbf{d}_r^{\text{in}} \in \mathbf{R}^{\text{in}}$ if and only if (3) holds for some k or (4) holds. To illustrate (3) and (4) from Table 2, let $\{v^1, \ldots, v^4\}$ be {"diabetes," "heart," "diabetes, heart," "diabetes, heart, MRSA"} and suppose $u_1 = 0.25$, $u_2 = 0.25$, $u_3 = 0.3$, $u_4 = 0.18$, $b = 0.6$, $c_1, \ldots, c_4 = 1$. Then, e.g., $\mathbf{d}_4^{\text{in}} = (N, N, N, H) \in \mathbf{R}_1^{\text{in}}$ because $P(v^1 \mid \mathbf{d}_4^{\text{in}}) = 0.5 > u_1 = 0.25$. Also notice that $\mathbf{d}_4^{\text{in}} \in \mathbf{R}_3^{\text{in}}$, illustrating the point that, although satisfying the criterion for being at risk for any confidential field value is enough to classify an input channel as risky, it may be that a given input channel is at risk for multiple values of the confidential field, and each of these sources of risk must be considered. Further, regardless of whether $\mathbf{d}_4^{\text{in}} \in \mathbf{R}_k^{\text{in}}$ for any k, $\mathbf{d}_4^{\text{in}} \in \mathbf{R}^{\text{in}}$ because $\sum_{k \in V'} P(v^k \mid \mathbf{d}_4^{\text{in}}) \cdot c_k = 1 > 0.6$.

5.2. The Safety of Output Channels Including Assignments

Here we establish conditions to determine whether a given potential output channel is safe and, if so, whether a given assignment of subjects to that output channel is likewise safe. The set of potential output channels is first limited to $\mathbf{S}^{\text{out}}$, which are those output channels that cannot be a priori rejected, i.e., output channels that cannot be determined to be infeasible independent of the x_{rw} variables. A simple condition is that $\mathbf{d}_w^{\text{out}} \in \mathbf{S}^{\text{out}}$ if and only if

$$\mathbf{d}_w^{\text{out}} \cap \mathbf{S}^{\text{in}} \neq \varnothing. \tag{5}$$

If (5) does not hold, the user knows that every subject in $\mathbf{d}_w^{\text{out}}$ comes from a risky input channel. The only uncertainties for the user would be which $\mathbf{d}_r^{\text{in}} \in (\mathbf{d}_w^{\text{out}} \cap \mathbf{R}^{\text{in}})$ correctly describes that subject, and therefore which element or elements of V cause that subject to be at risk. In addition, there are a posteriori conditions, based on the x_{rw} variables, that determine whether a given assignment of subjects to an output channel is safe. These conditions are analogous to (3) and (4) for input channels. Output channel w is *assignment safe* only if

$$P(v^k \mid \mathbf{d}_w^{\text{out}}) \leq u_k, \quad \text{all } k \in V, \tag{6}$$

or from the definition of $P(v^k \mid \mathbf{d}_w^{\text{out}})$ in (1),

$$\sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}}} (P(v^k \mid \mathbf{d}_r^{\text{in}}) - u_k) x_{rw} \leq 0, \quad \text{all } k \in V. \tag{7}$$

The second condition is the output channel analog of (4), namely

$$\sum_{j \in V'} c_j P(v^j \mid \mathbf{d}_w^{\text{out}}) \leq b, \tag{8}$$

which, from (1), is equivalent to

$$\sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}}} \left(b - \sum_{j \in V'} c_j P(v^j \mid \mathbf{d}_r^{\text{in}}) \right) x_{rw} \geq 0. \tag{9}$$

In summary, output channel w is "feasible," i.e., a priori and assignment safe, if and only if (5), (7), and (9) all hold.

We continue with the example of Figure 2, where $u_1 = 0.25,\ u_2 = 0.25,\ u_3 = 0.3,\ u_4 = 0.18,\ b = 0.6,$ $c_1, \ldots, c_4 = 1$, and focus on some possible output channels containing $\mathbf{d}_4^{\text{in}} \in \mathbf{R}^{\text{in}}$. The output channel $\{\mathbf{d}_1^{\text{in}}, \mathbf{d}_4^{\text{in}}\} \notin \mathbf{S}^{\text{out}}$ because it violates (5). On the other hand, consider $\{\mathbf{d}_4^{\text{in}}, \mathbf{d}_5^{\text{in}}\} \in \mathbf{S}^{\text{out}}$. If all subjects from $\mathbf{d}_4^{\text{in}}$ and $\mathbf{d}_5^{\text{in}}$ were assigned to that output channel, (7) would be violated for $k = 1$ and would therefore be assignment unsafe. Similarly consider $\{\mathbf{d}_3^{\text{in}}, \mathbf{d}_4^{\text{in}}\} \in \mathbf{S}^{\text{out}}$, and again suppose that all subjects in $\mathbf{d}_3^{\text{in}}$ and $\mathbf{d}_4^{\text{in}}$ were assigned to it. Then (7) and (9) would be satisfied so that the output channel is feasible. Note, however, if the data provider had set a lower value for b, say 0.45, then (9) would be violated.

6. A Linear Programming Model

6.1. The Objective Function

The ultimate objective of this technique is to provide the data recipients with a perturbed microdata set that has the highest possible utility and satisfies the security constraints set forth by the data provider. In this setting utility decreases whenever a subject, initially described by an input channel, is assigned to an output channel consisting of more than just that input channel. Furthermore, because the utility of the microdata set is reduced as subjects are associated with more common input channels, the "cost" of assigning a subject to output channel $\mathbf{d}_w^{\text{out}}$ is defined as $\psi_w^{\text{out}} := \sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}}} m_r^{\text{in}}$. This cost is measured as the number of subjects initially in the set of input channels that together comprise the output channel. This definition captures the "coarsening" of data that results from perturbation. It follows that $\psi_w^{\text{out}} = m$ if the output channel is composed of all input channels. That channel would provide the least possible amount of information about the subjects in the output channel. Thus, the data provider's goal is taken to be to minimize the total cost of the microdata set given by

$$\boldsymbol{\Psi} = \sum_{\mathbf{d}_w^{\text{out}} \in \mathbf{S}^{\text{out}}} \psi_w^{\text{out}} m_w^{\text{out}}, \tag{10}$$

subject to satisfaction of security constraints.

6.2. The Complete Model

Here we assume that $\mathbf{S}^{\text{out}}$ has been generated. Then the formulation is

$$\min \sum_{\mathbf{d}_w^{\text{out}} \in \mathbf{S}^{\text{out}}} \sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}}} \psi_w^{\text{out}} x_{rw} \tag{11}$$

$$\text{s.t.} \sum_{\mathbf{d}_w^{\text{out}} \ni \mathbf{d}_r^{\text{in}}} x_{rw} = m_r^{\text{in}}, \quad \mathbf{d}_r^{\text{in}} \in \mathbf{R}^{\text{in}} \tag{12}$$

$$\sum_{\mathbf{d}_w^{\text{out}} \ni \mathbf{d}_r^{\text{in}}} x_{rw} \le m_r^{\text{in}}, \quad \mathbf{d}_r^{\text{in}} \in \mathbf{S}^{\text{in}} \tag{13}$$

$$\sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}}} (P(v^k \mid \mathbf{d}_r^{\text{in}}) - u_k) x_{rw} \le 0, \quad \mathbf{d}_w^{\text{out}} \in \mathbf{S}^{\text{out}},\ k \in V \tag{14}$$

$$\sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}}} \left(b - \sum_{j \in V'} c_j P(v^j \mid \mathbf{d}_r^{\text{in}}) \right) x_{rw} \ge 0, \quad \mathbf{d}_w^{\text{out}} \in \mathbf{S}^{\text{out}} \tag{15}$$

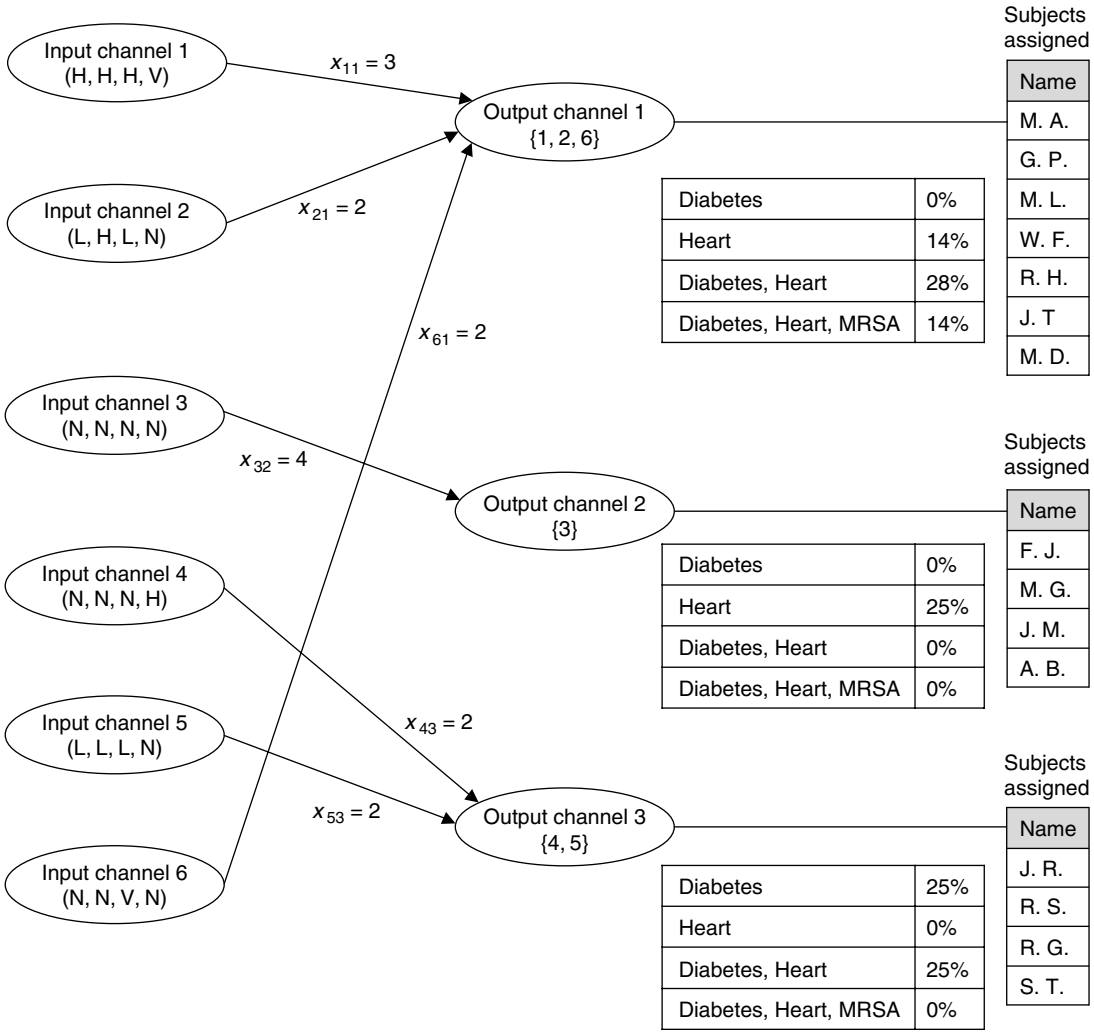
$$x_{rw} \ge 0, \quad \mathbf{d}_r^{\text{in}} \in \mathbf{d}_w^{\text{out}} \in \mathbf{S}^{\text{out}}. \tag{16}$$

Constraints (14) and (15) correspond to (7) and (9). Note that (13) is an inequality, because the remaining elements of the safe input channel are left assigned to themselves as safe, singleton output channels. It therefore follows that the costs of assigning subjects from the safe input channel $\mathbf{d}_r^{\text{in}}$ to the singleton output channel $\{\mathbf{d}_r^{\text{in}}\}$ are omitted from (11). Thus, (11) can be considered the degradation from the "least cost" solution of assignments only to singleton output channels. The formulation (11)–(16) has $\sum_{\mathbf{d}_w^{\text{out}} \in \mathbf{S}^{\text{out}}} |\mathbf{d}_w^{\text{out}}|$ decision variables and $|\mathbf{D}^{\text{in}}| + (|\mathbf{V}| + 1)|\mathbf{S}^{\text{out}}|$ constraints.

Solving the problem first presented in Table 1 (where $u_1 = 0.25,\ u_2 = 0.25,\ u_3 = 0.30,\ u_4 = 0.18,\ b = 0.6,\ c_1, \ldots, c_4 = 1$) using (11)–(16) yields the solution illustrated in Figure 5.

Notice that the subjects defined by input channel 1 and input channel 6 are used to provide protection for each other. Input channel 1 is at high risk for v^3 and v^4 and at low risk for v^1 and v^2. Input channel 6 has a nearly opposite risk profile, at low risk for v^2, v^3, and v^4 and high risk for v^1. By combining these two risky input channels, some risk mitigation takes place; however, the extent to which this can be done is obviously dependent on the values the data provider sets for b and c_k. In the case of the solution depicted in Figure 5, none of the input channels were "split" among two or more output channels, although that would not be the case in general, as is seen in Figure 6 and in the results of §7.

Figure 5 Optimal Solution to Initial Data Set of Table 1

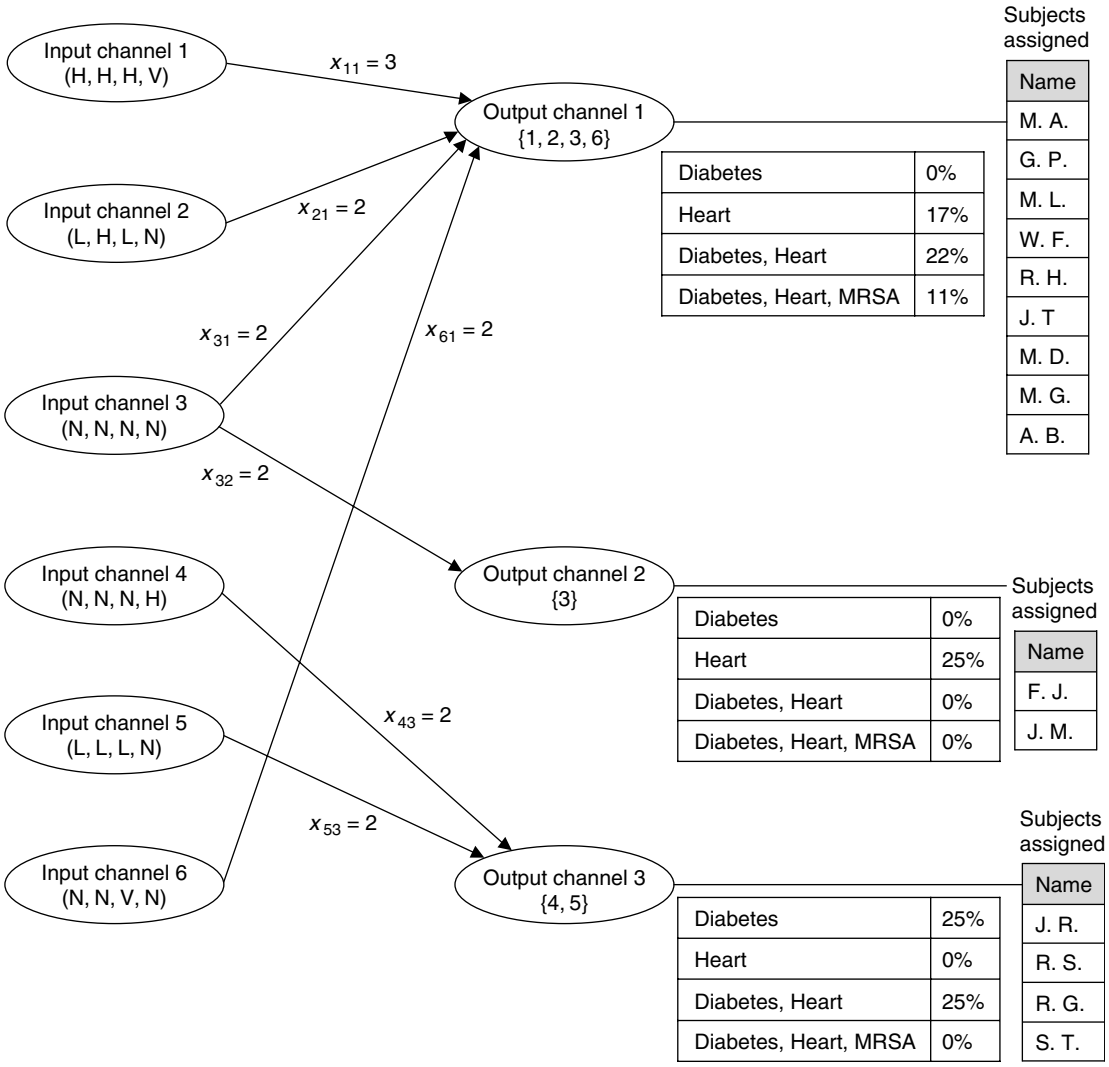


The optimal partial revelation solution illustrated in Figure 5 has greater utility than that depicted in Figure 4. Comparing the objective function values, the total cost of the strategy in Figure 4 is 117, while the cost associated with the strategy in Figure 5 is 65. This increase in utility can also be seen from the practical perspective of an IIM data user. Again, assume a public health worker wishes to identify and contact the subjects who are at risk for MRSA. From Figure 4, the public health worker would have received a list of nine subjects, three of whom were truly in the high-risk group and six spurious subjects. From Figure 5, the public health worker would have received a list of seven subjects, with the number of spurious subjects reduced to four.

To illustrate the role of the c_k parameters, assume that the data provider felt that the confidential field values {Diabetes, Heart} and {Diabetes, Heart, MRSA} both represent particularly sensitive combinations. The data provider could then set, for instance, $c_1 = c_2 = 1$ and $c_3 = c_4 = 1.25$. By increasing the "weight" that these confidential field values are assigned, the optimal solution would change from Figure 5 to Figure 6.

Notice that in the solution depicted in Figure 6, the effect of raising the values set for c_3 and c_4 was that the composition of output channel 1 changed and additional subjects from input channel 3 were assigned to output channel 1. This also resulted in a decrease in the probability that a user would associate

Figure 6 Optimal Solution Under Higher c_k's



with the likelihood that a subject described by output channel 1 has {Diabetes, Heart} or {Diabetes, Heart, MRSA}. However, changing the values of c_3 and c_4 will typically result in better solutions overall than simply lowering the values for u_3 and u_4. This is because lowering the values for u_3 and u_4 would impact the feasibility of every output channel, and lowering the values of c_3 and c_4 does not.

For example, consider an output channel with 100 subjects, 25 of whom have Diabetes and 28 of whom have {Diabetes, Heart}. Assume the risk parameters are initially set to values of $u_1 = 0.25$, $u_2 = 0.25$, $u_3 = 0.30$, $u_4 = 0.18$, $b = 0.6$, $c_1, \ldots, c_4 = 1$. Consider the impact of two approaches, lowering the value of u_3 and raising the value of c_3. If the value of c_3 is raised to 1.25, then the output channel is still feasible, because there is little risk elsewhere in the channel. However, if the value of u_3 is lowered to 0.22, then the output channel is no longer feasible. Lowering the values set for the u_k ignores the distribution of risk over the remaining values of the confidential field and can result in overly restrictive solutions. Also notice that Figure 6 depicts a solution in which the subjects initially described by one input channel (input channel 3) are spread across two output channels. The assignment of subjects from input

channel 3 to either output channel 1 or output channel 2 is done randomly; the x_{rw} simply dictate how many are to be assigned to each.

7. Computational Results

The computational experience follows the R-U confidentiality map framework (Duncan et al. 2003b). The goal of the computational experience is twofold: to gain insights into factors that impact the utility of the released data set and to provide useful guidelines that the data provider can use when establishing the values of important parameters such as $u_1, \ldots, u_K$, and b.

A 1 million-subject data set comprised of 200 input channels was created. Input channel cardinality was approximately uniformly distributed between 1,000 and 12,000 subjects. The confidential field contained two risky values (and implicitly a third value of "None"). The unconditional probability of each risky value was 0.1, and the conditional probability of each risky value within each input channel ranged from 0 to 0.5. The experimental design was to alter the proportion of subjects described by risky input channels from 0 to 0.20 in increments of 0.01. This design enabled us to evaluate situations where the amount of risk present in the original data set is extremely high, because in most realistic medical data sets it would be very unusual to have a situation in which this type of risk was more than 20% prevalent.

For each data set with a given proportion of subjects at risk, the linear program (LP) (11)–(16) was solved for different levels of risk tolerance. The computational experience was initiated by first considering a baseline level of risk tolerance where $u_1 = u_2 = 0.1$, $b = 0.15$, $c_1 = 0.5$, $c_2 = 1$. This baseline level of risk tolerance was then increased by a risk tolerance factor (RT) such that $u_1 = u_2 = 0.1 * \mathrm{RT}$ and $b = 0.15 * \mathrm{RT}$. RT was increased from a baseline value of 1 in increments of 0.2 to $\mathrm{RT} = 1.8$ (this corresponds to $u_1 = u_2 = 0.18$, $b = 0.27$).

To measure the utility of the released data set the normalized utility measure

$$\Omega = \sum_{\mathbf{d}_w^{\mathrm{out}} \in \mathbf{S}^{\mathrm{out}}} \sum_{\mathbf{d}_r^{\mathrm{in}} \in \mathbf{d}_w^{\mathrm{out}}} \psi_w^{\mathrm{out}} x_{rw} / m^2$$

was used to measure the decrease in the utility of the data as subjects were mapped from input channels to output channels. Its value will increase as the released data set becomes "coarsened" to a maximum value of 1. A value of 1 indicates that all subjects have been assigned to a single output channel that is comprised of every input channel in the original data set, and the data recipient essentially receives no information. A minimum value of 0 indicates that all subjects can be assigned to singleton output channels comprised of only the input channel that accurately describes each subject. In such a case the value of Ω is 0 because the objective function does not explicitly consider the costs associated with subjects that are assigned to singleton output channels. In such a situation the data recipient receives exact information, but that would only occur when no risk is present in the original data set.

The first step in the simulation process was to create a set of rules to generate the nonsingleton output channels. For this simulation the output channels were generated from the following rule set:

1. An output channel must contain at least one risky input channel and one safe input channel (this ensures that $\mathbf{d}_w^{\mathrm{out}} \in \mathbf{S}^{\mathrm{out}}$ for all w).

2. An output channel may not contain more than eight input channels. The only exception to this rule is the creation of an output channel comprised of all 200 input channels denoted by $\mathbf{d}_0^{\mathrm{out}}$.

3. The number of input channels contained in the output channels is biased toward output channels comprised of fewer input channels.

Singleton output channels were not generated because (a) it is, by definition, impossible for a subject from a risky channel to be assigned to a singleton output channel; and (b) any safe subjects allowed to remain unassigned by (13) can be thought of as being assigned to a singleton output channel comprised only of the input channel that originally described the subject. A total of 2,048 output channels were created based on the rule set outlined above.

Although not represented here, it is possible that output channels may be expected to satisfy other criteria possibly not involving risk. For example, the data provider may insist that they make sense from an application point of view and may specify a set of rules for discarding potential output channels as unreasonable. For instance, it may seem strange to be able to classify an individual as having either low

blood pressure or high blood pressure without the possibility of normal blood pressure.

Another possibility is that an output channel should not consist of an inordinate number of input channels. That is, it may be considered unacceptable to indicate that a given subject is described by one of five different input channels, because this may seem equivalent to simply saying that the subject is a member of the data set and nothing more. Thus, the task of enumeration of a reasonable candidate set $\mathbf{S}^{\text{out}}$ is likely to be computationally tractable in most real settings.

However, because this simulation did not enumerate all possible safe output channels, an analysis was conducted to evaluate the impact of restricting the set of output channels. To gain insight into the impact of the rule set on the quality of the solutions (that is limiting the number of output channels and the number of input channels that comprise each output channel), a problem was solved initially using just a single output channel and then solved repeatedly using increasingly large numbers of output channels. For the initial problem an input data set was chosen where the proportion of subjects at risk was 0.15 and where the data provider had set $u_1 = u_2 = 0.1$ and $b = 0.15$ to depict a situation where risk was fairly prevalent and risk tolerance was low. The problem was initially solved using a single output channel and then solved repeatedly with larger numbers of output channels.

To evaluate the impact of restricting the number of output channels under consideration, both the Ω and the proportion of unused output channels (output channels to which no subjects are assigned, i.e., $\sum_{\mathbf{d}_r^{\text{in}} \in \mathbf{S}^{\text{out}}} x_{rw} = 0$) were examined. The purpose of this analysis was to determine whether a point of diminishing return was present, after which the addition of more output channels did not result in better solutions but in an increase in unused output channels. Table 4 shows the results of the analysis.

In the case of limiting the number of output channels to 64, only 54 of the output channels were used. When the number of output channels was increased to 128, a slightly better solution was obtained that used 63 of the output channels. At 512 output channels a solution using 68 of the output channels yielded the best solution, which was not improved on with the addition of more output channels. So although Ω initially decreases dramatically, the improvement levels off when the number of output channels reaches 500. This suggests that imposing rules that restrict the number of output channels under consideration will not only help improve the "intuitive appeal" of the data, but also that imposing such rules is not expected to dramatically decrease the utility of the data set that is released.

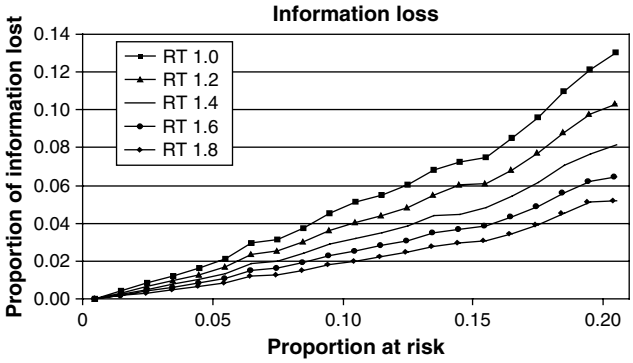
Table 4 Impact of the Number of Output Channels

Number of output channels	Ω	% Output channels unused
1	100.00	0.00
2	97.44	0.00
4	55.26	0.00
8	37.21	0.00
16	16.88	0.00
32	9.54	0.00
64	8.39	15.63
128	7.94	50.78
256	7.82	74.22
512	7.51	86.72
1,024	7.51	93.36
2,048	7.51	96.68

We then moved on to the large-scale simulation. The results of Figure 7 provide useful insights into the trade-off between risk tolerance and data utility and serve as an illustrative basis for recommendations that a data provider can employ to maximize the utility of a data set.

An interesting trend is that the rate of information loss increases as the proportion of risk increases, and that this trend is more pronounced when RT is low. Closer inspection shows that this effect is due

Figure 7 Simulation Results



to the assignment of subjects to the output channel $\mathbf{d}_0^{\text{out}}$. These represent very high-cost assignments and, while always feasible, cause significant erosion in the utility of the released data set. Although the assignment of subjects to $\mathbf{d}_0^{\text{out}}$ has a dramatic effect on the amount of information that is lost (because all information is lost on subjects assigned to that output channel), the assignment of subjects to $\mathbf{d}_0^{\text{out}}$ was also infrequent, even when the proportion of subjects at risk was high and risk tolerance was low. Table 5 shows the proportion of subjects assigned to $\mathbf{d}_0^{\text{out}}$.

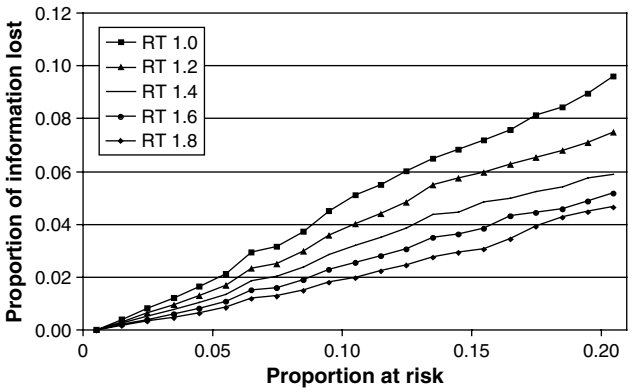
That the assignment of subjects to $\mathbf{d}_0^{\text{out}}$ is an infrequent occurance is important from the perspective of the practical utility of the technique outlined in this paper. Because the premise here is that individually identifiable microdata of high-risk subjects can be released in a safe and *useful* manner, it is important that, overwhelmingly, the data released is usable even when risk is pervasive and risk tolerance is low. Even in the highest risk scenario (lowest level of risk tolerance and largest proportion of subjects at risk), only about 2.2% of the subjects in the data set were assigned to $\mathbf{d}_0^{\text{out}}$. Figure 8 illustrates the information loss on the subjects that were not assigned to $\mathbf{d}_0^{\text{out}}$. A comparison of Figures 7 and 8 reveals that for more than 97% of the subjects, namely those not assigned to $\mathbf{d}_0^{\text{out}}$, the quality of the data released is significantly higher than that suggested in Figure 7. For example, in the highest risk scenario, the loss of information was reduced by more than 30% when the 2.2% of subjects in $\mathbf{d}_0^{\text{out}}$ were ignored.

Another key consideration for the data provider is the trade-off between the amount of security provided and data utility. Obviously, as the data provider becomes more risk averse and enforces lower values for $u_1, \ldots, u_K$ and b, the expected utility of the released data will decrease. It is therefore important

Table 5 **Proportion of Subjects Assigned to $\mathbf{d}_0^{\text{out}}$**

	Risk tolerance (RT)				
Proportion at risk	1.0	1.2	1.4	1.6	1.8
≤0.12	0	0	0	0	0
0.14	0.002	0.001	0	0	0
0.16	0.011	0.003	0.002	0	0
0.18	0.017	0.011	0.007	0.002	0
0.20	0.022	0.017	0.013	0.004	0.001

Figure 8 **Information Loss Excluding Subjects Assigned to $\mathbf{d}_0^{\text{out}}$**

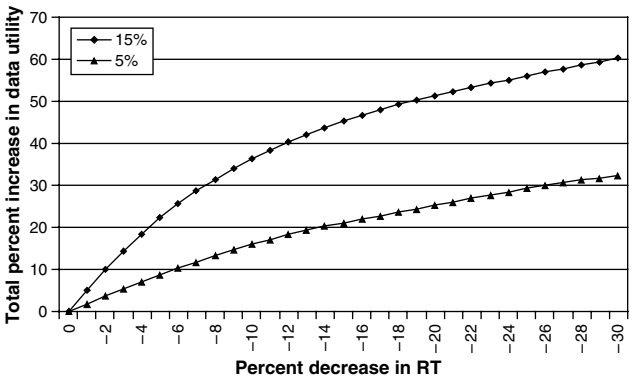


for the data provider to understand the details of the trade-off for the data set in question. Different data sets will exhibit different trends in terms of how much the data utility decreases in response to a decrease in risk tolerance, but the simulation data set provides some insights into how the data provider can approach the problem of striking a balance between the level of protection the subjects receive and the utility of the released data set.

To explore this trade-off, a simulation was conducted in which RT was initially set at 1.8 and then decreased in one percent increments. For each one percent decrease in RT we recorded both the total increase in data utility and the incremental increase in data utility. This was done for two data sets. In the first data set, 15% of the subjects were at risk, and in the second data set 5% of the subjects were at risk. The results for the total increase in data utility is depicted in Figure 9; the incremental increase in data utility for each successive one percent decrease in RT is shown in Figure 10.

Figures 9 and 10 illustrate the nonlinear trade-off between risk tolerance and data utility. In both cases, the data provider is motivated to avoid being too risk averse, because the gain in data utility is initially substantial as the data provider becomes more risk tolerant. The benefit is more pronounced for the data set where a greater proportion of subjects are at risk. These nonlinear trade-off patterns are likely to be common in real data sets because increasing risk tolerance can result in an improvement of the utility of the released data set in two manners. First, for higher levels of risk tolerance, fewer input channels

Figure 9 Total Increase in Data Utility



are expected to be considered risky. Second, for any input channel that does meet the criteria for being considered a threat, less work needs to be done to mitigate that threat.

Also notice that, in both cases, there is a point of diminishing return. For the data set where 15% of the subjects are at risk, reducing RT by more than 20% (from 1.8 to about 1.44) results in a situation where each additional one percent reduction in RT results in a less than one percent gain in data utility. For the data set where only 5% of the subjects are at risk, the point of diminishing return is reached earlier. In both cases the implications for data providers is clear. Significant gains in data utility can be obtained by moderate reductions in RT, and, at some point, the incremental gains in data utility do not justify additional reductions in RT. Because these trends may be more or less prevalent in different data sets, the data provider should have a thorough understanding of how different levels of risk tolerance impact data utility before making a final decision regarding the level of protection that will be provided.

The structure of the output channels, in terms of the number of input channels that comprise each output channel, was also examined. By examining the number of input channels that, on average, comprise the output channels, the data provider can gain some additional insights into the amount of "noise" that has been added to the data beyond the inclusion of spurious subjects to query results.

This is depicted in Figure 11 for all subjects, except those assigned to $\mathbf{d}_0^{\text{out}}$ and to singleton output channels. The rationale for excluding those assigned to $\mathbf{d}_0^{\text{out}}$ is that an output channel comprised of 200 input channels can dramatically skew the average, even when the number of subjects assigned to that channel is low.

Similarly, including safe subjects that have been assigned to singleton output channels can skew the average downward, especially when the proportion of at-risk subjects is low. Even when the proportion of subjects at risk is 0.2 and RT = 1, at-risk subjects are assigned to output channels that contain, on average, 4.5 input channels. This number drops significantly as risk tolerance increases. For the same proportion of subjects at risk, when RT = 1.8, at-risk subjects are assigned to output channels that contain, on average, 2.8 input channels. Considering that any at-risk subject will, by definition, be assigned to an output channel comprised of a minimum of two input channels (one safe channel and one risky channel), this represents "low noise" data.

All computations required for analysis were performed on a personal computer using CPLEX V.9.0.

Figure 10 Incremental Increase in Data Utility

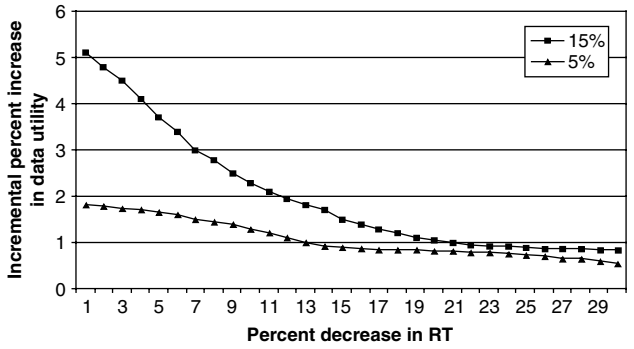
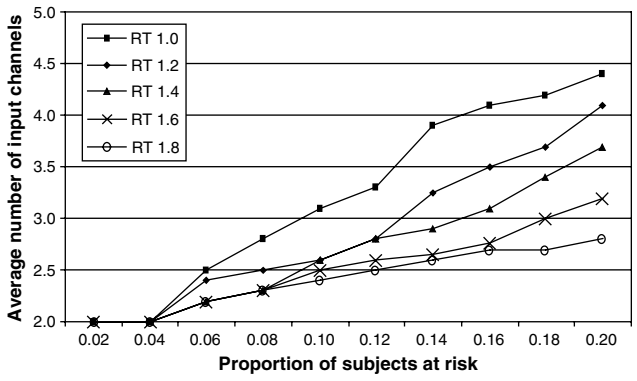


Figure 11 Average Number of Input Channels Ascribed to a Subject



Solutions for each scenario were obtained within 10 seconds of computational time. The ease of computational burden allows the data provider to extensively fine-tune the data release strategy to effectively trade off data security and data utility considerations.

8. Concluding Remarks

In this paper we consider security issues involved in releasing microdata with individual identifiers. The microdata provided to the users is inclusive in that a query of the output database will yield a set of subjects that includes all the subjects that would have been returned in the same query of the original, unperturbed database. The threat to the confidentiality of the subjects comes from the users possessing information that relates the microdata that is revealed to confidential information about the subjects. The general strategy we employ is to take the original data set in which some subjects are "safe" and some are at risk and transform it to a microdata set in which all subjects are safe. The problem of releasing as much data as possible, subject to the security constraints, is formulated as a linear program. Computational results suggest that the method is viable and that useful data can be released even when the level of risk in the input data set is high.

Natural extensions of this work would address the trade-offs between providing aggregate statistics and the quality of the disseminated microdata and additional considerations a data provider might place on which input channels can be combined to create the microdata set. The latter issue arises when, for example, the data provider wants to minimize alterations to some subset of the attributes in the microdata. In our work the revealed microdata do not contain confidential fields. Other techniques, such as perturbation, allow for the inclusion of these fields in the microdata, albeit in an altered format. Extensions of our approach to work within these settings is also a viable topic for future research.

Acknowledgments

The authors received support from TECI—The Treibick Electronic Commerce Initiative, Department of Operations and Information Management, University of Connecticut. The authors would also like to thank Dr. Manuel Nunez for his insightful comments and suggestions.

References

Adam, N. R., J. C. Wortmann. 1989. Security-control methods for statistical databases: A comparative study. *ACM Comput. Surveys* **21** 515–556.

Aggarwal, C. 2005. On *k*-anonymity and the curse of dimensionality. *Proc. 31st Internat. Conf. Very Large Databases* (*VLDB'05*), Trondheim, Norway.

Aggarwal, C., T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu. 2005. Approximation algorithms for *k*-anonymity. *J. Privacy Tech.*

Carvalho, F. D., N. Dellaert, M. S. Osorio. 1994. Statistical disclosure in two-dimensional tables: General tables. *J. Amer. Statist. Assoc.* **89** 1547–1557.

Causey, B. D., L. H. Cox, L. R. Ernst. 1985. Application of transportation thoery to statistical problems. *J. Amer. Statist. Assoc.* **80** 909.

CDC. 2005. HIPAA privacy rule and public health: Guidance from CDC and the U.S. Department of Health and Human Services. http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm.

Chowdhury, S. D., G. T. Duncan, R. Krishnan, S. F. Roehrig, S. Mukherjee. 1999. Disclosure detection in multivariate categorical databases: Auditing confidentiality protection through two new matrix operators. *Management Sci.* **45** 1710–1723.

Cox, L. H. 1980. Suppression methodology and statistical disclosure control. *J. Amer. Statist. Assoc.* **75** 377–385.

Cox, L. H. 1992. Solving confidentiality protection problems in tabulations using network optimization: A network model for cell suppression in U.S. economic censuses. R. Mokken, ed. *Internat. Sem. Statist. Confidentiality*. Eurostat, Dublin, Ireland, 229–245.

Dobra, A., S. E. Fienberg, M. Trottini. 2003. Assessing the risk of disclosure of confidential categorical data. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, M. West, eds. *Bayesian Statistics 7*. Oxford University Press, Oxford, UK, 125–144.

Domingo-Ferrer, J., J. M. Mateo-Sanz. 2002. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowledge Data Engrg.* **14**(1) 189–201.

Domingo-Ferrer, D., V. Torra. 2005. Ordinal, continuous, and heterogeneous *k*-anonymity through microaggregation. *Data Mining Knowledge Discovery* **11**(2) 195–212.

Duncan, G. T., S. Fienberg. 1999. Obtaining information while preserving privacy: A Markov perturbation method for tabular data. J. Domingo-Ferrer, ed. *Statist. Data Protection* (*SDP 1998*) *Proc.*, IDS Press, Amsterdam, The Netherlands.

Duncan, G., R. Krishnan, R. Padman, P. Reuther, S. Roehrig. 2003a. Exact and heuristic methods for cell suppression in multidimensional linked tables. Working paper, Carnegie Mellon University, Pittsburgh, PA.

Duncan, G. T., S. Keller-McNulty, S. L. Stokes. 2003b. Disclosure risk vs. data utility: The R-U confidentiality map. Technical Report 2003-6, Heinz School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA.

Fischetti, M., J. Salazar. 2001. Solving the cell suppression on tabular data with linear constraints. *Management Sci.* **47** 1008–1027.

Garfinkel, R., R. D. Gopal, P. B. Goes. 2002. Privacy protection of binary confidential data against deterministic, stochastic, and insider threat. *Management Sci.* **48** 749–764.

Gedik, B., L. Liu. 2005. A customizable *k*-anonymity model for protecting location privacy. *Proc. 25th Internat. Conf. Distributed Comput. Systems* (*IEEE ICDCS*), Columbus, OH.

Geurts, J. 1992. Heuristics for cell suppression in tables. Working paper, Central Bureau of Statistics, The Netherlands.

Gopal, R. D., P. B. Goes, R. Garfinkel. 1998. Interval protection of confidential information in a database. *INFORMS J. Comput.* **10** 309–322.

Health Insurance Portability and Accountability Act (HIPAA). 1996.

Hundepool, A., L. Willenborg. 1996. μ- and τ-*ARGUS*: Software for statistical disclosure control. *Proc. 3rd Internat. Seminar Statist. Confidentiality*, Bled, Slovenia.

Jiang, W., C. Clifton. 2005. Privacy-preserving distributed *k*-anonymity. *Proc. 19th Annual IFIP WG 11.3 Working Conf. Data Appl. Security*, Storrs, CT.

Kelly, J. P. 1990. Confidentiality protection in two and three-dimensional tables. Unpublished doctoral dissertation, University of Maryland, College Park, MD.

Li, X., S. Sarkar. 2006a. Privacy protection in data mining: A perturbation approach for categorical data. *Inform. Systems Res.* **17**(3) 254–270.

Li, X., S. Sarkar. 2006b. A tree-based perturbation approach for privacy-preserving data mining. *IEEE Trans. Knowledge Data Engrg.* **18**(9) 1278–1283.

Menon, S., S. Sarkar, S. Mukherjee. 2005. Maximizing accuracy of shared databases when concealing sensitive patterns. *Inform. Systems Res.* **16**(3) 256–270.

Meyerson, A., R. Williams. 2004. On the complexity of optimal *k*-anonymity. *Proc. 23rd ACM-SIGMOD-SIGACT-SIGART Sympos. Principles Database Systems*, Paris, France, 223–228.

Muralidhar, K., D. Batra, P. J. Kirs. 1995. Accessibility, security, and accuracy in statistical databases: The case for the multiplicative fixed data perturbation approach. *Management Sci.* **41** 1549–1564.

Newton, E., L. Sweeney, B. Malin. 2005. Preserving privacy by de-identifying facial images. *IEEE Trans. Knowledge Data Engrg.* **17**(2) 232–243.

Sweeney, L. 2002a. Guaranteeing anonymity when sharing medical data. *Proc. J. Amer. Medical Inform. Assoc.* Hanley & Belfus, Inc., Washington, D.C.

Sweeney, L. 2002b. Achieving *k*-anonymity privacy protection using generalization and suppression. *Internat. J. Uncertainty, Fuzziness Knowledge-Based Systems* **10**(5) 571–588.

United States General Accounting Office. 1999. Report to congressional requesters: Medical records privacy. (February).

Willenborg, L., A. De Waal. 1996. *Statistical Disclosure Control in Practice*. Springer-Verlag, New York.

Willenborg, L., A. Hundepool. 1998. ARGUS for statistical disclosure control. J. Domingo-Ferrer, ed. *Statist. Data Protection* (*SDP 1998*) *Proc.*, IDS Press, Amsterdam, The Netherlands.

Zayatz, L. 1992. Using linear programming methodology for disclosure avoidance purposes. Research report, Statistical Research Division, Bureau of the Census, Washington, D.C.