Measures of Disclosure Risk and Harm

Diane Lambert

AT&T Bell Laboratories

600 Mountain Avenue

Murray Hill, New Jersey 07974

Measures of Disclosure Risk and Harm

Diane Lambert

AT&T Bell Laboratories

Abstract

Disclosure is a difficult topic. Even the definition of disclosure depends on the context. Sometimes it is enough to violate anonymity. Sometimes sensitive information has to be revealed. Sometimes a disclosure is said to occur even though the information revealed is incorrect. This paper tries to untangle disclosure issues by differentiating between linking a respondent to a record and learning sensitive information from the linking. The extent to which a released record can be linked to a respondent determines disclosure risk; the information revealed when a respondent is linked to a released record determines disclosure harm. There can be harm even if the wrong record is identified or an incorrect sensitive value inferred. In this paper, measures of disclosure risk and harm that reflect what is learned about a respondent are studied, and some implications for data release policies are given.

This paper was written at the request of the Panel on Confidentiality and Data Access of the Committee on National Statistics of the National Research Council for a conference in 1991 entitled "Conference on Data Access Through Disclosure Limitation."

Keywords

Anonymity, confidentiality, disclosure threat, identification, linking, masked data, security

1. Introduction

Disclosure is a difficult topic. People even disagree about what constitutes a disclosure. Is it necessary to learn a sensitive attribute of a respondent or is it enough to identify an individual in a released file? Is there a disclosure when a sensitive attribute is learned but no record is associated with a particular individual? If first professions and then incomes are exchanged between records, it can be argued that no record belongs to the respondent but it seems dishonest to conclude that disclosures are thus impossible. Perhaps a disclosure occurs only when the new information is correct, perhaps not. If an intruder erroneously decides that a respondent is HIV positive or is ineligible for welfare, it may be little comfort that the information is wrong. Of course, the law may require one concept of disclosure and respondents may expect another. In any case, the agency cannot decide if confidentiality is protected without deciding what constitutes a disclosure.

Agencies often evaluate disclosure risk by considering just the data. Before release, the data are checked for extreme values and unique or unusual combinations of variables that could be used to identify respondents. But it is not the structure of the data alone that determines whether disclosure is likely. The individual or organization that receives or surreptitiously accesses the data plays an equally important role in determining whether disclosure is likely. The legitimate researcher with an ongoing need for agency data has little incentive to try to identify individual respondents. But an immature computer hacker challenged by the difficulty of re-identifying a released record, a disgruntled employee bent on discrediting an agency, or a company seeking information about a competitor's plans or an employee's behavior may have more incentive, and more resources, to re-identify a record. Whether such people or companies exist is a question beyond the scope of this paper. But what is undeniable is that the risk of disclosure depends on who gains access to the data and what strategies they use to compromise confidentiality. The person, or group of persons, who attempts a disclosure has been called a data user, intruder, attacker, data spy, data snoop or snooper. None of these terms is perfect, but one is needed. As in Duncan and Lambert (1989), the term intruder is used in this paper.

One could argue that models of disclosure are hopeless because the issues are too complex and the intruder too mysterious. Instead, this paper argues that models of disclosure are indispensable. At the least, they force definitions and assumptions to be stated explicitly. And, when the assumptions are realistic, models of disclosure can lead to practical measures of disclosure risk and harm.

In the disclosure limitation framework proposed in Duncan and Lambert (1986, 1987, 1989), the intruder plays a central role. This is important, but the models of Spruill (1982, 1983, 1984), Paass (1985, 1988), Bethlehem, Keller, and Pannekoek (1990), and Keller-McNulty and Unger (1990) have other strengths of their own. This paper expands the disclosure limitation framework to include some of the strengths in these other papers and to derive measures of harm. Section 2 distinguishes re-identifications or violations of anonymity from disclosures of sensitive information and separates true disclosures from perceived disclosures. Section 3 reviews the basic premise of the disclosure limitation framework for re-identification, which is that the risk of perceived re-identification depends on how strongly the intruder links a target respondent to a released record, regardless of whether the link is correct or incorrect. This risk depends on properties of the data, the respondents, and the intruder. Section 4 computes perceived disclosure for two extended examples, and shows that some common beliefs about disclosure are correct and others are incorrect. Section 5 proposes a practical measure of the risk of disclosing a true identity to an intruder and relates it to a strategy that some agencies have used to assess how well confidentiality is protected. Section 6 proposes measures of harm that take into account the record re-identified and the sensitive attribute inferred. The paper concludes with some implications for data access policies.

2. What is a Disclosure?

An agency has records on many people in a file. Each record has attributes, such as age, location, marital status and profession, that are useful for identification but are not usually sensitive. Bethelehem, Keller and Pannekoek (1990) call these *key variables*. Other attributes on the records, such as diseases, debts, and credit rating, may be sensitive. Suppose a sample of the records is released, with obvious identifiers such as name and address eliminated, some attributes such as marital status left intact, and other key and sensitive attributes modified to preserve confidentiality. For example, incomes might be truncated, professions grouped more coarsely, and ages on pairs of records swapped. Further, some attributes on some records might be missing or imputed rather than observed. What could the released data disclose?

There are two major types of disclosure about individuals. In an *identity disclosure*, or *identification*, a respondent is linked to a particular record in a released file. Identification, sometimes called re-identification, is equivalent to inadvertent release of an identifiable record. With microdata, only respondents whose records are released can be correctly re-identified. Identifications are also possible from tabular data and inquiries about groups, however. If there is only one female black dentist in an area and the right sequence of queries reveals that she is in the database, then an identification occurs. Even if the intruder learns nothing sensitive from the identification, the re-identification itself may compromise the security of the data file.

An *attribute disclosure* occurs when the intruder believes something new has been learned about the respondent. An attribute disclosure may occur with or without an identification. For example, suppose all union plumbers in Chicago earn the same wage and the Department of Labor releases the average wage of union plumbers in Chicago as part of a table. Then the release tells all about the wage of any union plumber in Chicago, although no record is identified with a respondent. Similarly, the intruder may narrow the list of possible target records to two with nearly the same value of a sensitive attribute. Then the attribute is disclosed although the target record is not located. Or two records may be averaged so the released record belongs to no one. Yet the debt on the averaged record may disclose something about the debt carried by the targeted individual. The agency must decide whether attribute disclosures without identifications are important.

This paper, following recent trends in disclosure analysis, considers only disclosures that involve re-identifications. Attribute disclosures without re-identifications are not considered (but see Duncan and Lambert 1986 and Skinner 1992). Attribute disclosures that result from re-identification are considered to the extent that they harm the respondent. That is, in this paper, the *risk of disclosure* is the risk of re-identifying a released record and the *harm from disclosure* depends on what is learned from the identification.

Attribute disclosures that do not involve identification are often ignored, as they are here (*e.g.*, Bethlehem, Keller, and Pannekoek 1990, Greenberg and Voshell 1990, and Keller-McNulty and Unger 1990). This restriction can be challenged, however. For example, it assumes that all intruders first look for the record that is most likely to be correct and then take information about the targeted attribute from that record. Intruders with other strategies are ignored.

Many papers equate disclosure with revealing the true identity or attribute of a respondent (e.g., Blien, Wirth and Muller 1992). This is semantically reasonable, but it dismisses some disclosures that cause harm. In contrast, the disclosure limitation model of Duncan and Lambert (1986, 1987, 1989) does not separate true and false disclosures, since what matters there is what the intruder believes has been disclosed. But then harm is difficult to measure, because false inferences may have different consequences from true ones. Keller-McNulty and Unger (1990) take an intermediate position. They consider only correct identifications but allow incorrect attribute inferences to follow from a correct identification. This paper goes one step further and includes true and false re-identifications and true and false attribute disclosures. Correct and incorrect inferences can be distinguished if desired (as happens with measures of harm), but they need not be. This leaves to others the question of whether only true disclosures should be considered.

In this paper, when only correct inferences are to be prevented the terms *true identification* and *true attribute disclosure* are used, as in Keller-McNulty and Unger (1990). When the objective is to prevent the intruder from believing that there is a disclosure, regardless of whether the information taken from the released data is correct, the terms *perceived identification* and *perceived attribute disclosure* are used.

Finally, all examples in this paper pertain to static data releases, but there are no conceptual barriers to applying the framework to dynamic databases, as Duncan and Mukherjee (1991, 1992a, 1992b) show.

3. The Risk of Perceived Identification

The basic premise in the disclosure limitation framework is simple but controversial. An agency protects confidentiality by discouraging the intruder rather than by encouraging false re-identifications or incorrect inferences about sensitive attributes. This stance has several rationales. An agency's mandate is not to mislead. False disclosures may harm a respondent as much as true disclosures. False disclosures may be impossible to deny and may erode survey participation as much as a true disclosure would. Thus, disclosure is limited only to the extent that the intruder is discouraged from making any inferences, correct or incorrect, about a particular *target* respondent.

Evaluating whether an intruder is discouraged requires thinking like an intruder, a rational intruder who makes optimal decisions based on the perceived probability of success and perceived value of the information to be gained. That is, the agency must consider disclosure from the perspective of an intruder.

The intruder sees the problem as follows. An agency holds N records in a file $\mathbf{Z}$ and releases a random sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of n masked records with k attributes each. (If the intruder does not know that the target individual is included in the agency's file, then $\mathbf{Z}$ refers to the population and N refers to the number of individuals in $\mathbf{Z}$ rather than the number of individuals interviewed.) Masking suppresses attributes in $\boldsymbol{Z}$, adds random noise, truncates outliers, or swaps values of an attribute between records. Knowing this, which, if any, record in the released file should be linked to the target respondent?

The intruder has two options. One is to decide that one of the released records belongs to the target respondent. That is, the intruder could link the i^{th} released record $\mathbf{x}_i$ to the target record $\mathbf{Y}$. Or, the intruder could decide not to link any released record to $\mathbf{Y}$, perhaps because none of the released records is close enough to what the intruder expects for $\mathbf{Y}$ or perhaps because too many released records are close to what the intruder expects for $\mathbf{Y}$. The decision not to link is here called the *null link*. Not surprisingly, the rational intruder chooses the link (non-null or null) believed most likely to be correct whenever any incorrect choice incurs the same positive loss and a correct link incurs no loss. (See Duncan and Lambert (1989) for details.)

More precisely, let p_i be the intruder's probability that the i^{th} released record in $\boldsymbol{X}$ is the target's. Then $1 - \sum_{i=1}^{n} p_i$ is the intruder's probability that the target record has not been released. When the perceived probability that the target record has not been released exceeds

$\max_{1 \le i \le n} p_i$, the intruder chooses not to link the respondent to any released record, as the agency prefers. But when $\max_{1 \le i \le n} p_i$ is large enough, the intruder does link the target to a released record. Because $\max_{1 \le i \le n} p_i$ measures how much the data reveal to the intruder about which released record, if any, is the target's, it is here called the *risk of perceived identification* or, more specifically, the *risk of perceived re-identification*, for the target record $\mathbf{Y}$.

Note that the risk of perceived identification is not defined in terms of the intruder's expected loss or the agency's or respondent's expected loss, but, less formally, in terms of the seriousness of the threat posed by the intruder. Further, the risk of perceived re-identification depends only on the intruder's posterior probability that the i^{th} released record is the correct one. It does not depend on what, if anything, is revealed to the intruder after a link is made. Also note that if the intruder is a group, then p_i is the intruding group's consensus probability. The model would have to be more complicated if the group did not have a consensus probability. See, for example, Mokken, Kooiman, Pannekoek,and Willenborg (1992).

There are various ways that the risks of perceived disclosure for different respondents can be combined to define a risk of perceived identification for an entire source file $\mathbf{X}$. Focusing on the most easily identified respondent gives the worst case or *pessimistic* risk

$$\begin{aligned} D(\boldsymbol{X}) &= \max_{1 \le j \le N} \max_{1 \le i \le n} P[i^{th} \textit{ released record is } j^{th} \textit{ respondent's record} \mid \boldsymbol{X}] \\ &= \max_{1 \le j \le N} \max_{1 \le i \le n} P[\mathbf{x}_i \textit{ is } j^{th} \textit{ respondent's record} \mid \boldsymbol{X}]. \end{aligned}$$

The pessimistic measure protects against the intruder who is looking for the easiest record to identify. It does not matter whether the intruder chooses a record to match back to a respondent, perhaps because the record has "interesting" attributes, or chooses a respondent to match to a record. The pessimistic measure also guards against the intruder who is trying to re-identify more than one record, since

$$D(\boldsymbol{X}) \ge \max_{1 \le i,\, j \le n} P[\mathbf{x}_i \textit{ is respondent } 1\textit{'s record} \text{ and } \mathbf{x}_j \textit{ is respondent } 2\textit{'s record} \mid \boldsymbol{X}].$$

Risks for different respondents can also be combined by averaging instead of maximizing:

$$D_{average}(\mathbf{X}) = N^{-1} \sum_{j=1}^{N} \max_{1 \le i \le n} P[\mathbf{x}_i \textit{ is } j^{th} \textit{ respondent's record} \mid \boldsymbol{X}],$$

or by summing instead of maximizing:

$$D_{total}(\mathbf{X}) = \sum_{j=1}^{N} \max_{1 \le i \le n} P[\mathbf{x}_i \textit{ is } j^{th} \textit{ respondent's record} \mid \boldsymbol{X}].$$

Alternatively, the total risk of perceived identification for $\mathbf{X}$ can be defined to be the number of respondents for whom the risk of perceived identification is above a threshold τ:

$$D_\tau(\mathbf{X}) \;=\; \sum_{j=1}^{N} \#\{1 \le j \le N\colon \max_{1 \le i \le n} P[\mathbf{x}_i \textit{ is } j^{th} \textit{ respondent's record} \mid \boldsymbol{X}] \ge \tau\}.$$

How the risks for individual respondents are combined is less important than recognizing that the risks depend on the intruder's perceptions.

These measures of the risk of disclosure for $\mathbf{X}$ apply to a wider class of intruders than those who minimize expected loss and assign the same loss to choosing an incorrect record and not linking at all. Less is revealed, and these measures are conservative, when the intruder feels choosing a wrong record is worse than not linking. Less is also disclosed when the intruder is discouraged from choosing the most probable link, even if the decision process is too complicated to model by loss functions. (See Section 4.2.4 for such an example.) They are not completely general, however. Intruders who seek to discredit an agency by announcing a link has been made, regardless of how unlikely the link is, have a much larger loss for a null link than for an incorrect link and are not covered. Of course, in practice a released file may be attacked by several different groups of intruders, and combining their separate risks is difficult. For an approach to this problem, see Mokken, Kooiman, Pannekoek and Willenborg (1992).

4. Modeling the Intruder

Thinking about the intruder's probability p_i that the i^{th} released record is the target is difficult but not impossible. Elaborations of two examples taken from Duncan and Lambert (1989) suggest how the intruder's perceptions and the risk of identification can be evaluated.

4.1. Example 1: A Knowledgeable Intruder and a Population of Two Records

Often, the intruder knows something about the target, but not precisely how its record will appear if released. The release may be inaccurate, the respondent may not have been forthcoming when interviewed, information may change with time, or the data may have been altered to protect against disclosure. The question is, how will approximate knowledge be used?

Suppose there is one, continuous attribute and the intruder is willing to make judgments about the masked version $M(\mathbf{Y})$ of the target $\mathbf{Y}$. For example, the intruder may believe that $M(\mathbf{Y})$ is sure to be within 100% of a number m_1 and probably within 30% of m_1. Suppose that after a series of such judgments the intruder concludes that $M(\mathbf{Y})$ is lognormal(μ_1 , σ_1). Then the intruder believes that if the target record is released, it is equally likely to be above or below e^{μ_1}, and with probability 0.95 it is within a factor of $e^{\pm 2\sigma_1}$ of e^{μ_1}. Whether the imprecision is caused by lack of prior knowledge or masking or both does not matter. To be specific, take $\mu_1 = 0$ and $\sigma_1 = 1$.

Information about other respondents cannot be ignored because it may help to identify the target. Suppose the intruder's beliefs about the other respondent's released attribute $M(\mathbf{Y}_2)$ are modeled by a lognormal(2, 1) distribution. Then the intruder expects the target to be released as 1.6 and the other respondent as 12.2. The actual released data, though, are $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2) = (7, 20)$. One of these values must be the target's, and the only reasonable decision is that the record with 7 is. The question is, how sure is the intruder of the link?

After the data are released,

$$p_1 = P[M(\mathbf{Y}_1) = 7 \mid \mathbf{X} = (7, 20)]$$

$$= \frac{f_1(7) f_2(20)}{f_1(7) f_2(20) + f_1(20) f_2(7)} = 0.89,$$

where f_i is the lognormal(μ_i, 1) density. That is, with probability 0.89 the intruder links the target to $\mathbf{x}_1$, even though 7 is much larger than the intruder would have predicted. Plainly, prior information about the position of the target relative to the other respondent helps the intruder. Note that $P[\mathbf{Y}_1 = 7] = P[\mathbf{Y}_2 = 20]$, so the pessimistic risk of perceived identification for this source file is

$$D(\mathbf{X}) = \max_{1 \le j \le 2,\ 1 \le i \le 2} P[M(\mathbf{Y}_j) = x_i \mid \mathbf{X} = (7, 20)] = 0.89.$$

Now suppose that the agency releases only one record from this population of two and the released record contains $\mathbf{x}_1 = 7$. Then

$$p_1 = P[\mathbf{Y}_1 \textit{ sampled} \text{ and } M(\mathbf{Y}_1) = 7 \mid \mathbf{X} = 7]$$

$$= \frac{P[\mathbf{Y}_1 \textit{ sampled} \text{ and } M(\mathbf{Y}_1) = 7]}{P[\mathbf{Y}_1 \textit{ sampled} \text{ and } M(\mathbf{Y}_1) = 7] + P[\mathbf{Y}_2 \textit{ sampled} \text{ and } M(\mathbf{Y}_2) = 7]}$$

$$= \frac{.5 f_1(7)}{.5 f_1(7) + .5 f_2(7)} = 0.13\ .$$

The probability that the released record is not the target's, i.e. that the released record belongs to the other respondent, is $1 - 0.13 = 0.87$. Then the perceived risk of identification for the other respondent is 0.87, and the pessimistic risk of a perceived identification for $\mathbf{X}$ is $\max(0.13,\ 0.87) = 0.87$. The risk of perceived identification does not depend on whether the released record belongs to the target or the other respondent, *i.e.*, on whether the linking is correct. This is sensible because the spy's decision to act and compromise confidentiality can be based only on what is perceived to be a link, not on what is in fact a true link.

This simple example illustrates three major points. First, the intruder need not explicitly ''undo'' the masking to re-identify the target. The masking merely changes the intruder's probability distributions, perhaps just increasing their imprecision. Second, the rational intruder considers all respondents in $\boldsymbol{Z}$. Here 7 was an unlikely value for the target when only one record was released, but not when two records were released and the other record was 20. Third, the agency cannot keep the risk of identification at zero for all intruders, but it can identify a class of intruders who pose little risk. For example, the risk may be low when the intruder's prior distributions on the masked records are lognormal with imprecisions above a given level and medians closer than a given amount.

4.2. Example 2: A Knowledgeable Intruder and a Sample of n Records from N

Suppose the intruder believes that if the i^{th} record out of N is released, it will appear as $M(\mathbf{Y}_i)$ with density $f_i(\cdot)$. Then the intruder's probability that the n^{th} released record belongs to the target $\mathbf{Y}_1$ is

$$p_n = P[\mathbf{Y}_1 \textit{ is sampled} \text{ and } M(\mathbf{Y}_1) = \mathbf{x}_n \mid \boldsymbol{X}] \tag{1}$$

$$= \frac{P[\mathbf{x}_n \textit{ sampled from } f_1 \text{ and } \mathbf{x}_1, \ldots, \mathbf{x}_{n-1} \textit{ sampled from } f_2, \ldots, f_N]}{P[\mathbf{x}_1, \ldots, \mathbf{x}_n \textit{ sampled from } f_1, \ldots, f_N]}$$

The denominator is the probability of the released data given the intruder's beliefs about all N respondents (or all N individuals in the population if the intruder does not know who was interviewed). This simple formula leads to several conclusions about disclosure.

4.2.1. Respondents who are not unique in the population are protected.

When the intruder cannot *a priori* distinguish respondents, all f_i's are the same, and the risk of re-identification reduces to $1/N$, which is typically negligible. Likewise, if an intruder cannot *a priori* distinguish two respondents, say $\mathbf{Y}_1$ and $\mathbf{Y}_2$, the risk of identification for either respondent is at most 0.5. The proof for $N = 3$ and $n = 2$ gives the flavor of the general argument. There,

$$p_1 = \frac{f_1(\mathbf{x}_1)[f_2(\mathbf{x}_2) + f_3(\mathbf{x}_2)]}{f_1(\mathbf{x}_1)[f_2(\mathbf{x}_2) + f_3(\mathbf{x}_2)] + f_2(\mathbf{x}_1)[f_1(\mathbf{x}_2) + f_3(\mathbf{x}_2)] + f_3(\mathbf{x}_1)[f_1(\mathbf{x}_2) + f_2(\mathbf{x}_2)]}$$

Since $f_1(\cdot) = f_2(\cdot)$, $p_1 \leq 0.5$, with strict inequality unless $f_3(\boldsymbol{x}_1) = 0$ or $\mathbf{x}_2$ is impossible under f_1. Even with precise prior information about the population, *e.g.* with f_i concentrated at one value, undifferentiated respondents cannot be re-identified confidently. Consequently, if the intruder knows that there are no unique records in a population, there is little risk of perceived re-identification for any respondent.

Unfortunately, even agency files with little detail may have many unique records. Bethlehem, Keller and Pannekoek (1990) describe a file concerning 23,485 households in the Netherlands with 2 parents and 2 children. Each record consists of the sexes of the children and the ages to the nearest year of all household members. Although there are only 6 attributes, and 2 of these are binary, 68% of the households are unique.

4.2.2. Confidentiality is protected if the released file has no unique records.

Sometimes, the risk of re-identification is assumed to depend on the number of unique records in the released file rather than the number of unique records in the source file (see, for example, Bethlehem, Keller and Pannekoek 1990, Greenberg and Voshell 1990, Keller and Bethlehem 1992, and Greenberg and Zayatz 1992). Formula 1 shows why this is correct. No matter who the target respondent is, two identical released records $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same probability of being the target. Consequently, if no unique masked record is released, the pessimistic risk and average risk of perceived identification are at most 0.5 if at most two records are identical and less if more released records are identical. That is, if $\mathbf{X}$ has no unique records, the intruder is not certain that any link is correct. Unfortunately, substantial aggregation, rounding and grouping may be needed to avoid having unique records in released samples.

4.2.3. Unknown respondents may be re-identifiable.

Bethlehem, Keller and Pannekoek (1990) claim that if an intruder knows nothing about a respondent, that respondent cannot be re-identified. They write, "If someone has no information about a specific individual, identification and thus disclosure is impossible. Hence, the risk of disclosure depends on the nature and amount of *a priori* available knowledge." This statement is reasonable, but it can be false. For example, suppose that $N = 2,\ n = 1$ and one attribute is released. The masked version of $\mathbf{Y}_2$ is assumed to be normal(0, 1). Little is known about the target $\mathbf{Y}_1$, so the intruder assigns its masked record a uniform(-4, 4) distribution. The one record released contains the value -2.25. Then

$$p_1 \;=\; \frac{1/8}{1/8 \,+\, f_2(-2.25)} \;=\; 0.80 \;.$$

That is, the intruder need not know much about the target respondent if much is known about the other respondents in the population. Such examples may not be pathological if the respondents are business establishments, the target is a new company about which little is known, and the other respondents are old companies.

4.2.4. Sampling by itself need not protect confidentiality.

Intuitively, releasing a small fraction of the records in an agency's file should protect against disclosure. But consider the following scenario. There are $N = 100$ records in the agency's file and 10 randomly chosen masked records are released, each having one attribute. The intruder, who believes the target $\mathbf{Y}_1$ has the smallest value in the population, models uncertainty about $\mathbf{Y}_1$ with a lognormal(0, 0.5) distribution. The true values of the 99 other respondents are modeled as a lognormal(2, 0.5) random sample. (Prior distributions must be specified for all respondents, but not all respondents need to be distinguished.) The intruder assumes that the agency's masking effectively multiplies the data by an independent random lognormal(0, 0.5) inflation factor, giving a median inflation factor of 1 and a 95% chance that the inflation factor is between 0.38 and 2.66. As a result, the intruder models the masked values of $\mathbf{Y}_2, \ldots, \mathbf{Y}_{100}$ as a lognormal(2, 1) random sample and the masked value of $\mathbf{Y}_1$ as a lognormal(0, 1) random variable. Note that $f_i(\cdot)$, the density assigned to $\mathbf{Y}_i$, combines uncertainty about the values in the source file with the uncertainty the agency introduces by masking.

The agency releases 10 masked records, say 0.05, 0.14, 1.5, 2.4, 3.2, 3.8, 4.6, 8.7, 10.3 and 10.7. Since all densities except the target's are the same, p_i can be re-written as

$$p_i = \frac{P[\mathbf{x}_i \textit{ is } \mathbf{Y}_1\textit{'s}]\, f_1(\mathbf{x}_i)\, P[9 \textit{ records from } \mathbf{Y}_2, \ldots, \mathbf{Y}_{100} | \mathbf{Y}_1 \textit{ sampled}] \prod_{j \neq i} f_2(\mathbf{x}_j)}{\sum_{j=1}^{10} P[\mathbf{x}_j \textit{ from } \mathbf{Y}_1 \text{ and } \textit{other } 9 \textit{ records from } \mathbf{Y}_2, \ldots, \mathbf{Y}_{100}] + P[\mathbf{x}_1, \ldots, \mathbf{x}_{10} \textit{ from } \mathbf{Y}_2, \ldots, \mathbf{Y}_{100}]}$$

$$= \frac{\frac{1}{100} f_1(\mathbf{x}_i) \prod_{j \neq i} f_2(\mathbf{x}_j)}{\sum_{j=1}^{10} \frac{1}{100} f_1(\mathbf{x}_j) \prod_{k \neq j} f_2(\mathbf{x}_j) + \frac{\begin{bmatrix} 99 \\ 10 \end{bmatrix}}{\begin{bmatrix} 100 \\ 10 \end{bmatrix}} \prod_{k=1}^{10} f_2(\mathbf{x}_k)}$$

$$= \frac{f_1(\mathbf{x}_i)/f_2(\mathbf{x}_i)}{\sum_{j=1}^{10} f_1(\mathbf{x}_j)/f_2(\mathbf{x}_j) + 90},$$

where $f_1(\cdot)$ is the lognormal(0, 1) density and $f_2(\cdot)$ the lognormal(2, 1) density. Evaluating p_i shows that the intruder assigns the records with 0.05 and 0.14 to the target with probabilities 0.86 and 0.11 respectively. Each of the other 8 records has at most 0.001 probability, and the probability that $\mathbf{Y}_1$ has not been released is 0.026. None of the other respondents can be distinguished, so their probabilities of identification are all less than 1/99. Therefore, the pessimistic risk of perceived re-identification is $D(\boldsymbol{X}) = 0.86$. That is, there is a substantial risk of perceived re-identification for at least one respondent even though only 10 records with one attribute each have been released from a population of 100 records.

Of course, less is revealed, and the risk of perceived disclosure is conservative, if the intruder does not link to the most probable record. For instance, the intruder might choose not to link at all if a p_i as large as 0.86 were likely in samples without the target record. Here, for example, 10,000 lognormal(2, 1) random samples of size 10, *i.e.*, samples of 10 non-target records, simulated in S (Becker, Chambers and Wilks 1988) gave an estimated mean $\max_{1 \leq i \leq n} p_i$ of 0.05 and estimated upper 0.95 and 0.99 quantiles of $\max_{1 \leq i \leq n} p_i$ of 0.19 and 0.41 respectively. Hence, if the intruder's perceptions about the 99 other respondents are correct, it is unlikely that a probability as high as 0.86 would occur if the target record were not released. If the simulation had shown that 0.86 was likely by chance, the intruder could decide not to link and $D(\boldsymbol{X})$ would then be conservative.

4.2.5. Summary

There are four major points. First, the risk of identification is less than 0.5 if no unique masked records are released or if it is common knowledge that there are no unique respondents in the population. Second, an intruder with little prior information about the masked target record, other than a range of possible values, may be able to re-identify the target record if enough is known about the other respondents. (This point was also made in a report published in the United States by the Subcommittee on Disclosure Avoidance Techniques (1978), which was a group of representatives from several United States government agencies and others interested in disclosure issues. Third, releasing just a small sample of records may not protect all respondents. Fourth, less is disclosed, and the risk of perceived identification is overstated, if the intruder can be dissuaded from linking to the most probable record.

Finally, it is important to note that the model does not specify a safe level for the probability of perceived identification, and that even low levels of identification risk may sometimes be unacceptable. A referee gives the following example. Suppose an intruder knows that 5 percent of the people with a sensitive attribute $Y = 1$ will respond to a direct mail campaign. Also suppose that the intruder is able to obtain a list of people, 10 percent of whom have $Y = 1$ and the remainder of whom do not. Then the intruder could mount a direct mail campaign with a response rate of .5 percent, which is very high. Thus, the risk of identification and the risk of correctly inferring $Y = 1$ are low, but the intruder has profited from the released data and the agency's reputation may be severely damaged.

5. The Risk of True Identification

It is easy to tell whether an intruder can re-identify a respondent correctly. For example, the intruder in Section 4.2.4 believes that the best match is the record with $\mathbf{x}_1 = 0.05$. If it is, and the intruder judges that a probability of 0.86 is high enough to act on, there is a true identification.

The agency cannot control the intruder's perceptions and actions once the data are released. All it can do is count the number of true identifications for an intruder with a given set of beliefs about the target and source file. A reasonable measure of the *risk of true identification*, then, is simply the fraction of released records (or number of released records) that an intruder can correctly re-identify. (A slight variant would be to weight each correct re-identification by the probability that the intruder assigns to the re-identified respondent.)

In Section 4.2.4, for example, the intruder believes that 99 of the 100 respondents have probability at most 1/99 of being re-identified. If 1/99 is too low a probability for the intruder to act on, none of these 99 records will be correctly re-identified. If the remaining record is correctly re-identified, the risk of true identification is 1/100; if not, the risk is zero. Of course, a different intruder would lead to a different numerical risk of true identification.

Spruill (1982, 1983, 1984) proposed the first practical measure of disclosure. In its simplest form, it is computed in two steps. First, find the distance between each masked record and each source record. Next, find the fraction of masked records that are closer to their unmasked parent than they are to any of the other $N - 1$ source records. Weighted square distances and weighted absolute distances were suggested, although more complicated distances may be used in practice. The fraction of released records that are closer to their parent record than to any other source record has been called the risk of disclosure, but to emphasize the distinction with the risk of true identification, it is called the *risk of matching* here.

The risk of true identification is, loosely speaking, also based on distance. To see this, suppose that the sampling fraction is so low and the agency file so large that sampling records for release without replacement is nearly the same as sampling them with replacement. Then equation (1) becomes

$$p_n = \frac{P[\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} \text{ } \textit{sampled from } f_2, \ldots, f_N \text{ and } \mathbf{x}_n \text{ } \textit{sampled from } f_1]}{P[\mathbf{x}_1, \ldots, \mathbf{x}_n \text{ } \textit{sampled from } f_1, \ldots, f_N]}$$

$$\approx \frac{f_1(\mathbf{x}_n) \prod_{i=1}^{n-1} P[\mathbf{x}_i \text{ } \textit{from } f_2 \textit{ or } f_3 \textit{ or } \cdots \; f_N]}{\prod_{i=1}^{n} P[\mathbf{x}_i \text{ } \textit{from } f_1 \textit{ or } f_2 \textit{ or } \cdots \; f_N]}$$

$$= \frac{f_1(\mathbf{x}_n) \prod_{i=1}^{n-1} \sum_{j=2}^{N} f_j(\mathbf{x}_i)}{\prod_{i=1}^{n} \sum_{j=1}^{N} f_j(\mathbf{x}_i)} .$$

It follows that $\mathbf{x}_1$ is a more likely for the target $\mathbf{Y}_1$ than $\mathbf{x}_n$ is, *i.e.* $p_1 > p_n$, if

$$\frac{f_1(\mathbf{x}_1)}{\sum_{j=2}^{N} f_j(\mathbf{x}_1)} > \frac{f_1(\mathbf{x}_n)}{\sum_{j=2}^{N} f_j(\mathbf{x}_n)}. \tag{2}$$

When each f_i is a normal($\mathbf{y}_i$, $\mathbf{\Sigma}_i$) density, so that the intruder's beliefs are centered at the values $\mathbf{y}_i$ on the source records, the right side of inequality (2) reduces to

$$\frac{e^{-\frac{1}{2}(\mathbf{x}_n - \mathbf{y}_1)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mathbf{y}_1)}}{\sum_{i=2}^{N} e^{-\frac{1}{2}(\mathbf{x}_n - \mathbf{y}_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mathbf{y}_i)}}.$$

This term is large when $\mathbf{x}_n$ is close to the target record $\mathbf{y}_1$ or far from the other source records $\mathbf{y}_2, \ldots, \mathbf{y}_N$. That is, the risk of true identification, like the risk of matching, depends on a notion of closeness, albeit an unusual distance that depends on all N respondents. The weights Σ_i used here resemble the weights in the risk of matching, but here the weights are motivated by the intruder's uncertainty about the masked respondents rather than by the variance of the $\mathbf{y}_i$'s. Note that absolute distance arises if double exponential densities replace the normal densities in inequality (2).

The risk of true identification and the risk of true matching can both be interpreted in terms of distance measures, but they are fundamentally different, as a simple example shows. Suppose the source file has the 10 records shown in Table 1. (These values were generated in S as a lognormal(2, 0.5) random sample of size 10). The two records 15.0 and 30.0 are chosen for release, and masked to 32 and 35 respectively. The intruder, who knows the values on the source records, is told that the released records have been multiplied by a lognormal(0, 0.5) random factor and takes f_i to be a lognormal(y_i, 0.5) density. The probabilities in Table 1 are then computed to determine which, if any, released records can be re-identified.

TABLE 1

The intruder's probability that the released record in the row comes from the source record in the column. Unknown to the intruder, 32 is the masked version of 15.0 and 35 is the masked version of 30.0.

	source									
	9.8	10.8	14.1	14.6	14.7	15.0	30.0	40.7	47.1	53.2
released										
32	0.016	0.024	0.065	0.072	0.074	0.078	0.202	0.183	0.156	0.130
35	0.010	0.017	0.048	0.054	0.056	0.059	0.199	0.205	0.188	0.164

From Table 1, the intruder finds that the source record 15.0 is more likely to have been released as the masked record 32 than as the masked record 35, which is a correct inference. The probability that the link is correct, however, is only 0.078 which may be too low for the intruder to act on. If so, the source record 15.0 is not re-identified. The source record 30.0 is more likely to have been released as 32 than as 35, which is incorrect, and thus it does not count as a true disclosure. Therefore, the risk of true identification is zero.

The risk of matching is not zero, however. The released record 35 is closest to the source record 30.0, and it is a true match. The released record 32 is also closest to the source record 30, but that is an incorrect match. The risk of matching is thus 1/2.

Loosely speaking, the risk of matching compares distances across rows of Table 1 and the risk of true identification compares distances down columns. The risk of matching fixes the released record and finds the closest respondent. Because the risk of true identification is based on the intruder's strategy, it fixes the respondent, finds the closest released record, and counts the correct matches that are probable. If the intruder acts only when the probability of the link is at least 0.5, *i.e.*, only when the link is more likely to be correct than incorrect, then only released records that are more likely to have come from the target respondent than any other respondent can be counted as true matches. (This is true because the probabilities for any released record sum to one across respondents.) In this sense, using the intruder's perspective leads to a measure that naturally considers all released records and all source records simultaneously. The risk of true matching, however, ignores the possibility that some other released record might be a better match for the target source record and does not discard correct matches as too improbable.

6. Disclosure Harm

6.1. Harm to Whom?

Harm could refer to any undesirable consequence from a disclosure, perceived or true. The agency itself, for example, could suffer from a disclosure. Respondents might become reluctant to participate in its surveys without time-consuming and costly personal interviews. Congress might restrict the agency's ability to share sample frames with other agencies, driving up its costs but not its budget. The agency's employees might have to dedicate more time and resources to confidentiality studies. Legitimate researchers could be harmed by a disclosure, too. Future released microdata might be too heavily masked to support legitimate analyses. Then the public could be hurt because public policy debate would be less well-informed.

These consequences are serious and real, and should not be underestimated. But they are not studied here. If the intent is to protect confidentiality of respondents, it is more important to measure harm to the respondent than it is to measure harm to the agency, researcher, or public policy.

Besides, harm to public policy follows harm to the researcher, which follows from harm to the agency, which in turn is most likely to follow from a respondent being harmed by a re-identification. Each step in the chain away from the respondent becomes more difficult to quantify, and each depends on the first re-identification. Thus, in this paper, harm from disclosure means harm to a respondent whose released record has been re-identified, or has been perceived to be re-identified.

Sometimes, harm is restricted to consequences that follow from true disclosures. Bethlehem, Keller and Pannekoek (1990) unequivocally state that "If the person is not in the sample, no harm can be done." Other authors just as unequivocally claim that perceived, untrue disclosures can be more damaging than true disclosures. Ladd (1989) describes disastrous consequences of incorrect re-identification of criminal records. Less dramatically, the Privacy Protection Study Commission Report of 1977 replaced the earlier Health, Education and Welfare notion of harm to the respondent with the notion of fair and accurate record keeping (Ware 1980). This suggests that the worst consequences can be avoided by preventing untrue disclosures. Since people strongly disagree about whether true or false disclosures are worse, the measures of harm defined here allow either position.

6.2. A Scenario

The intruder wants to know whether a target respondent, Y_{10}, is HIV positive, has used crack, or has some other characteristic that is not public knowledge. What harm could there be if the agency releases a masked microdata file $\mathbf{X}$ that contains, among other variables, the (binary) attribute of interest for a sample of respondents, one of whom might be the target?

Since the focus is on harm from re-identification, the intruder must first find the target's record in $\mathbf{X}$. Suppose the intruder links the target to the released record $\mathbf{x}_1$. The knowledgeable intruder realizes that the sensitive attribute, say x_{11}, on the re-identified record could be wrong. For example, masking may have changed an occurrence ($x_{11} = 1$) into a non-occurrence ($x_{11} = 0$) or conversely. Suppose the intruder believes that

$$x_{11} = \begin{cases} Y_{11} & \textit{with probability} \;\; q \\ 1 - Y_{11} & \textit{with probability} \;\; (1 - q) \end{cases}$$

where Y_{11} indicates whether the target has the sensitive attribute and q is independent of the other attributes on the record. That is, $P[x_{11} = 1 | Y_{11} \text{ and } \mathbf{x}] = P[x_{11} = 1 | Y_{11}]$. How does this affect the intruder's inference about Y_{11}?

Let $\mathbf{x}_{-11}$ be $(x_{12}, ..., x_{1k})$; *i.e.* the re-identified record with the sensitive attribute removed, and $\mathbf{x}_{-1} = (\mathbf{x}_{-11,} \; \mathbf{x}_2, ..., \mathbf{x}_n)$. Then if $x_{11} = 1$,

$$P[Y_{11} = 1 | \mathbf{x}] = P[Y_{11} = 1 | x_{11} \text{ and } \mathbf{x}_{-1}]$$

$$= \frac{P[x_{11} = 1 | Y_{11} = 1, \mathbf{x}_{-1}]\, P[Y_{11} = 1 | \mathbf{x}_{-1}]}{P[x_{11} = 1 | Y_{11} = 1, \mathbf{x}_{-1}]\, P[Y_{11} = 1 | \mathbf{x}_{-1}] + P[x_{11} = 1 | Y_{11} = 0, \mathbf{x}_{-1}]\, P[Y_{11} = 0 | \mathbf{x}_{-1}]}$$

$$= \frac{qP[Y_{11} = 1 | \mathbf{x}_{-1}]}{qP[Y_{11} = 1 | \mathbf{x}_{-1}] + (1 - q)P[Y_{11} = 0 | \mathbf{x}_{-1}]}. \qquad (3)$$

Similarly, if $x_{11} = 0$,

$$P[Y_{11} = 1 | \mathbf{x}] = \frac{(1 - q)P[Y_{11} = 1 | \mathbf{x}_{-11}]}{(1 - q)P[Y_{11} = 1 | \mathbf{x}_{-11}] + qP[Y_{11} = 0 | \mathbf{x}_{-11}]}.$$

If $P[Y_{11} = 1 | \mathbf{x}]$ is close enough to one, the intruder decides that the target belongs to the sensitive category. If $P[Y_{11} = 1 | \mathbf{x}]$ is close enough to zero, the intruder decides that the target does not belong to the sensitive category. If $q = 1$, then $P[Y_{11} = x_{11} | \mathbf{x}] = 1$ and the intruder reads the sensitive category from the re-identified record. If $q = 0$, the intruder believes the opposite of what is on the re-identified record to be true. For $0 < q < 1$, the intruder may infer the value on the re-identified record, the opposite value, or no value if $P[Y_{11} = 1 | \mathbf{x}]$ is sufficiently close to 0.5. That is, re-identification does not inevitably lead to disclosing a sensitive attribute.

The sophisticated intruder uses all the released data to infer $P[Y_{11} = x_{11} | \mathbf{x}]$. Suppose that the probability that a respondent has the sensitive attribute depends on the other $k - 1$ observed attributes through the logistic relationship

$$\log\left[\frac{P[Y_{i1} = 1 | (x_{i2}, \ldots, x_{ik})]}{P[Y_{i1} = 0 | (x_{i2}, \ldots, x_{ik})]}\right] = \beta_0 + \sum_{j=2}^{k} \beta_j x_{ij}.$$

The true Y_{i1}'s are not observed, but $\boldsymbol{\beta}$ can be estimated by maximizing the log-likelihood in terms of the recorded x_{i1}'s, which is

$$\sum_{i=1}^{n} \log\left[P[x_{i1} = 1 | \boldsymbol{x}_{-i1}]\right] = \sum_{i=1}^{n} \log\left[P[x_{i1} = 1, Y_{i1} = 1 | \boldsymbol{x}_{-i1}] + P[x_{i1} = 1, Y_{i1} = 0 | \boldsymbol{x}_{-i1}]\right]$$

$$= \sum_{i=1}^{n} \log\left[q\, P[Y_{i1} = 1 | \boldsymbol{x}_{-i1}] + (1 - q)\, P[Y_{i1} = 0 | \boldsymbol{x}_{-i1}]\right]$$

Estimates of $\boldsymbol{\beta}$ that accommodate measurement error in the released attributes $x_{i2}, \ldots, x_{ik}$ are also possible (see, for example, Carroll *et al.* (1984)).

Once $\boldsymbol{\beta}$ is estimated, the probability that $Y_{11} = 1$ given the entire released file and the information on the re-identified record is estimated by substituting

$$\hat{P}[Y_{11} = 1 | \boldsymbol{x}_1] = \frac{e^{\hat{\beta}_0 + \sum_{j=2}^{k} \hat{\beta}_j x_{1j}}}{1 + e^{\hat{\beta}_0 + \sum_{j=2}^{k} \hat{\beta}_j x_{1j}}}$$

into equation (3).

For inferences about continuous sensitive attributes from masked data, see Tendick and Matloff (1987, 1993) and Sullivan and Fuller (1989). These references, and the example above, assume that a intruder constructs a point estimate of the target's sensitive value. A Bayesian intruder, however, might construct a posterior probability distribution instead.

6.3. Measures of Harm

Once the data are released, the agency and respondent cannot affect the intruder's perceptions. Given a particular intruder, the harm that the respondent experiences is determined by the intruder; it is not random. (Unless, of course, the intruder randomly chooses the sensitive attribute by flipping a coin.) Assessing harm amounts to enumerating all the possible decisions that the intruder can make and evaluating the consequences to the respondent of each decision. For a binary sensitive attribute, the harm to respondent 1 because of an inference about Y_{11} after the intruder perceives the target record to be re-identified is

$$H(Y_{11}, \mathbf{X}) = \begin{cases} 0 & \text{if } \textit{record not reidentified} \\ c_{FN} & \text{if } \textit{reidentification incorrect} \text{ and } Y_{11} \textit{ not inferred} \\ c_{TN} & \text{if } \textit{reidentification correct} \text{ and } Y_{11} \textit{ not inferred} \\ c_{FF} & \text{if } \textit{reidentification incorrect} \text{ and } Y_{11} \textit{ inferred incorrectly}, \\ c_{FT} & \text{if } \textit{reidentification incorrect} \text{ and } Y_{11} \textit{ inferred correctly} \\ c_{TF} & \text{if } \textit{reidentification correct} \text{ and } Y_{11} \textit{ inferred incorrectly} \\ c_{TT} & \text{if } \textit{reidentification correct} \text{ and } Y_{11} \textit{ inferred correctly} \end{cases}$$

and the total harm caused by $\mathbf{X}$ is

$$H_{total}(Y_1, \mathbf{X}) = \sum_{i=1}^{N} H(Y_{i1}, \mathbf{X}).$$

Any *consequence* c_{ij} can be set to zero. For example, suppose an intruder re-identifies the wrong record but does nothing further. Some respondents may feel this is an intolerable invasion of privacy and assign a large, positive c_{FN}. Others may be unconcerned and assign $c_{FN} = 0$. Similarly, some c_{ij} 's may be equal. If the consequences depend on which value of the attribute is inferred but not on the correctness of the re-identification, $c_{FF} = c_{TF}$ and $c_{FT} = c_{TT}$. A model cannot make these kinds of judgments. At best it can be rich enough to accommodate a variety of perspectives.

Simple measures of harm may be the most useful because they are easily understood. If so, the best measures of harm for binary attributes count "inferences", just as the most popular measures of the risk of re-identification count identifications. For example, suppose no respondent wants an intruder to decide that $Y_{i1} = 1$, and all respondents agree on the damage from this inference. Then the total harm is just the number of respondents that the intruder assigns to the sensitive category. Or if the agency focuses on correctly re-identified records, the total harm is the number of respondents that the intruder correctly re-identifies and then (correctly or incorrectly) assigns to the sensitive category. Thus, the total true disclosure risk is the number of records that the intruder correctly re-identifies and the total disclosure harm is the number of correctly re-identified records that the intruder assigns to the sensitive category.

For a continuous sensitive attribute, an appropriate measure of harm is

$$H(Y_{11},\, \mathbf{X}) \;=\; \begin{cases} 0 & \text{if } \mathit{record\ not\ reidentified} \\ c_{FN} & \text{if } \mathit{incorrect\ record\ re-identified} \text{ and } Y_{11}\ \mathit{not\ inferred} \\ c_{TN} & \text{if } \mathit{correct\ record\ re-identified} \text{ and } Y_{11}\ \mathit{not\ inferred} \\ c_F(y_{11}, \hat{y}_{11}) & \text{if } \mathit{reidentification\ incorrect} \text{ and } \hat{y}_{11}\ \mathit{inferred} \\ c_T(y_{11}, \hat{y}_{11}) & \text{if } \mathit{reidentification\ correct} \text{ and } \hat{y}_{11}\ \mathit{inferred} \end{cases} .$$

The two functions c_F and c_T may be the same if the correctness of the re-identification is unimportant, and they may depend on how close the inferred attribute is to the true value. For example, if y_{11} is the target's average charge card balance, a respondent may be harmed only by overestimates. If so, harm might be measured by

$$c_F(y_{i1}, \hat{y}_{i1}) \;=\; c_T(y_{i1}, \hat{y}_{i1}) \;=\; \begin{cases} 0 & \text{if } \hat{y}_{i1} < y_{i1} \\ (\hat{y}_{i1} - y_{i1})^2 & \text{if } \hat{y}_{i1} \ge y_{i1} \end{cases} .$$

If the respondent is penalized only if y_{i1} exceeds a threshold *T*, a better measure of harm is

$$c_F(y_{i1}, \hat{y}_{i1}) \;=\; c_T(y_{i1}, \hat{y}_{i1}) \;=\; \begin{cases} 0 & \text{if } \hat{y}_{i1} \le T \\ L & \text{if } \hat{y}_{i1} > T \end{cases} .$$

Counting inferences may not measure harm from continuous attributes adequately. If respondents agree that the amount by which an inferred continuous attribute exceeds the true value is important, then a better measure of total harm is

$$H_{total}(Y_{1,}\mathbf{X}) = \sum_{i=1}^{N} max(0,\ \hat{y}_{i1} - y_{i1}).$$

Finally, a Bayesian intruder computes a posterior density $p(y_{11})$ for the sensitive attribute rather than a point estimate. The harm from a Bayesian intruder inferring a binary attribute for a target respondent is

$$H_{Bayes}(Y_{11},\ \mathbf{X}) = \begin{cases} 0 & \text{if } record\ not\ re-identified \\ p(y_{11} = 1)\,c_{F1} + p(y_{11} = 0)\,c_{F0} & \text{if } incorrect\ record\ re-identified \\ p(y_{11} = 1)\,c_{T1} + p(y_{11} = 0)\,c_{T0} & \text{if } correct\ record\ re-identified \end{cases}.$$

The harm from a Bayesian intruder inferring a continuous attribute for a target respondent is

$$H_{Bayes}(Y_{11},\ \mathbf{X}) = \begin{cases} 0 & \text{if } record\ not\ re-identified \\ \int p(y)\,c_F(y_{11},y)\,dy & \text{if } wrong\ record\ re-identified \\ \int p(y)\,c_T(y_{11},y)\,dy & \text{if } correct\ record\ re-identified \end{cases}.$$

Since the harm from a Bayesian intruder is a weighted average of the harm from point estimates, there may be no advantage to computing H_{Bayes}.

6.4. Assigning Numerical Values to Disclosure Harm

There are studies that try to assign values to public goods such as clean air or undeveloped forest land. Values have been assigned using surveys and mock auctions (e.g., Brookshire and Coursey 1987, Brookshire, Thayer, Schulze, and D'Arge 1982) and focus groups and surveys (e.g., Desvousges and Smith 1982, Desvousges and Frey 1989). People are asked, usually indirectly, how much they are willing to pay to improve the current state or how much they need to be compensated to accept a worse state. Studies must be carefully constructed to elicit accurate information, or else participants exaggerate the value of the public good (Brookshire and Coursey (1987). The most successful studies have elicited market information about something that affects the participants directly, such as planting more trees in an existing park. Often, supporting material, such as landscape drawings of the park with different densities of trees, are needed to show the different levels of the public good. If the situation can be described fully and a reasonable facsimile of a market set up, well-designed studies provide information about how people value public goods.

Confidentiality may be more difficult to evaluate than the environment, though. People may not know the current level of disclosure. They may be unaware that agencies sometimes share data and that there are many private databases for marketing, medical insurance, credit reporting, and other purposes. People may not know that researchers may be given masked individual records, or that empirical studies (e.g. Paass 1985, 1988) have shown that a large fraction of masked records in very large databases can be re-identified. If the current status is not appreciated, it will be difficult to elicit the consequences to disclosures. It is also seems difficult to simulate a market that will force people to reveal how much they would pay for confidentiality. Nonetheless, the literature on valuing public goods may give some insight into measuring how the public evaluates confidentiality.

7. Discussion

Once data are released, it is the intruder, and not the structure of the data alone, that controls disclosure. When the intruder is sure enough that a released record belongs to a respondent, there is a re-identification. It may be incorrect, but the intruder perceives there to be a re-identification. Plainly, the risk of perceived disclosure and the risk of true disclosure cannot be measured without considering the seriousness of the threat posed by the intruder's strategy. Likewise, the harm that follows from a re-identification depends on the attributes, if any, that the intruder infers about the target, and so harm cannot be measured without considering the strategy that the intruder uses to infer sensitive attributes. But once the intruder's strategy is modeled, disclosure risk and harm can be evaluated, as shown in this paper.

All the agency can do to reduce disclosure risk or disclosure harm is to mask the data before release or carefully select the individuals and organizations that are given the data, or both. Unfortunately, the model developed here and in Duncan and Lambert (1986, 1989) implies that masking and releasing only a subset of records does not necessarily protect against disclosure. This is not the fault of the model. Empirical studies by Paass (1985, 1988) give the same conclusion. The empirical study by Blien, Wirth, and Muller (1992) also shows that the perceived risk of identification is nonzero even if the intruder requires the released record to match the prior information exactly and the released data are noisy. Masking may lower the risk of true re-identification, but it may also lead to false re-identifications and false inferences about attributes. The fact that inferred attributes may be wrong may be little comfort to the respondent whose record is re-identified.

Masking also complicates data analysis. For example, a good masking procedure either preserves the first two moments of joint distributions or allows them to be recovered. But many interesting analyses do not involve the mean and covariance matrix of the joint distribution. For example, the relationship between $\mathbf{x}$ and $\mathbf{y}$ for people in the 90th percentile of variable $\mathbf{z}$ involves more than the first two moments of $(\mathbf{x}, \mathbf{y}, \mathbf{z})$. Nonlinear relationships or relationships with binary data also require more than two moments. An agency cannot be expected to predict and minimize all the effects of masking on all the analyses of interest. Nor is it reasonable to expect the data analyst to describe how the data will be analyzed before the data are obtained so that the agency can verify that the conclusions will be the same for the masked data as they would have been for the original data. Future masking techniques may preserve more general features of the data, but for now data masked enough to preserve confidentiality can be a challenge to analyze appropriately.

It does seem reasonable to put some of the burden for protecting confidentiality on the researcher, however. Institutions and researchers have to abide by all sorts of conditions in experiments involving humans. The experience in those and other areas ought to provide some guidance on protecting respondents in agency databases from unscrupulous intruders. This would not necessarily remove the need for some masking, but it might reduce the need for heroic masking that severely limits the usefulness of the data.

References

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988). The New S Language. California: Wadsworth & Brooks/Cole.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38-45.

Blien, U., Wirth, H., and Muler, M. (1992). Disclosure Risk for Microdata Stemming from Official Statistics. Statistica Neerlandica, 46, 69-82.

Brookshire, D. S. and Coursey, D. L. (1987). Measuring the Value of a Common Good: An Empirical Comparison of Elicitation Procedures. The American Economic Review, 77, 554-566.

Brookshire, D. S., Thayer, M. A., Schulze, W. D. and D'Arge, R. C. (1982). Valuing Public Goods: A Comparison of Survey and Hedonic Approaches. The American Economic Review, 72, 165-177.

Carroll, R. J., Spiegelman, C. H., Lan, K. K., Bently, K. T., and Abbott, R. D. (1984). On Errors-in-Variables for Binary Regression Models. Biometrika, 71, 19-25.

DeGroot, M. H. (1970). Optimal Statistical Decisions. New York: McGraw-Hill Book Company.

Desvousges, W. H. and Smith, V. K. (1988). Focus Groups and Risk Communication: The "Science" of Listening to Data. Risk Analysis, 8, 479-484.

Desvousges, W. H. and Frey, J. H. (1989). Integrating Focus Groups and Surveys: Examples from Environmental Risk Studies. Journal of Official Statistics, 5, 349-363.

Duncan, G. T. and Lambert, D. (1986). Disclosure-Limited Data Dissemination (with discussion). Journal of the American Statistical Association, 81, 10-28.

Duncan, G. T. and Lambert, D. (1987). The Risk of Disclosure for Microdata. Proceedings of the Third Annual Research Conference of the U.S. Bureau of Census.

Duncan, G. T. and Lambert, D. (1989). The Risk of Disclosure for Microdata. Journal of Business and Economic Statistics, 7, 207-217.

Duncan, G. T. and Mukherjee, S. (1991). Microdata Disclosure Limitation in Statistical

Databases: Query Size and Random Sample Query Control. Proceedings of the 1991 IEEE Computer Security Symposium on Research in Security and Privacy. Oakland, California, May 20-22, 278-287.

Duncan, G. T. and Mukherjee, S. (1992a). Disclosure Limitation Using Autocorrelated Noise. Proceedings of the 1992 IFIP Conference on Computer Security. August 23. Vancouver, BC.

Duncan, G. T. and Mukherjee, S. (1992b). Confidentiality Protection in Statistical Databases: A Disclosure Limitation Approach. Proceedings of the International Seminar on Statistical Confidentiality, organized by Eurostat and the International Statistical Institute, September, 1992, Dublin, Ireland.

Greenberg, B. V., and Voshell, L. (1990). Relating Risk of Disclosure for Microdata and Geographic Area Size. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Greenberg, B. V. and Zayatz, L. V. (1990). Strategies for Measuring Risk in Public Use Microdata Files. Statistica Neerlandica, 46, 33-48.

Keller, W. J., and Bethlehem, J. G. (1992). Disclosure Protection of Microdata. Statistica Neerlandica, 46, 5-20.

Keller-McNulty, S. and Unger, E. A. (1990). The Deterrent Value of Natural Change in a Statistical Database. Proceedings of the Statistical Computing Section, American Statistical Association, 15-23.

Ladd, J. (1989). Computers and Moral Responsibility: A Framework for an Ethical Analysis. The Information Web, Ethical and Social Implications of Computer Networking, ed. C. C. Gould. London: Westview Press.

Mokken, R.J., Kooiman, P., Pannekoek, J., and Willenborg, L. C. R. J. (1992). Disclosure Risks for Microdata. Statistica Neerlandica, 46, 49-68.

Paass, G. (1985). Disclosure Risk and Disclosure Avoidance for Microdata. Journal of Business and Economic Statistics, 6, 487-500.

Paass, G. (1988). Disclosure Risk and Disclosure Avoidance for Microdata. Presented to the Conference on Access to Public Data, Social Science Research Council, November 1985, Washington, D.C.

Skinner, C. J., (1992). On Identification Disclosure and Prediction Disclosure for Microdata. Statistica Neerlandica, 46, 21-32.

Spruill, N. L. (1982). Measures of Confidentiality. Statistics of Income and Related Administrative Record Research: 1982, Washington, D.C.: U.S. Department of the Treasury, Internal Revenue Service, Statistics of Income Division, 131-136.

Spruill, N. L. (1983). The Confidentiality and Analytic Usefulness of Masked Business Microdata. Proceedings of the Section on Survey Research Methods, American Statistical Association, 602-607.

Spruill, N. L. (1984). Protecting Confidentiality of Business Microdata by Masking, Public Research Institute, Alexandria, Va.

Subcommittee on Disclosure Avoidance Techniques (Federal Committee on Statistical Methodology) (1978). Statistical Working Paper 2, Federal Statistical Policy and Standards. Washington, DC: U.S. Department of Commerce.

Sullivan, G. and Fuller, W.A. (1989). The Use of Measurement Error to Avoid Disclosure. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Tendick, P. and Matloff, N. S. (1987). Recent Results on the Noise Addition Method for Database Security. Proceedings of the Section on Survey Research Methods, American Statistical Association, 406-409.

Tendick, P. and Matloff, N. S. (1993). A Modified Random Perturbation Method for Database Security. Transactions on Database Systems, Association for Computing Machinery, to appear.

Ware, W. (1980). Privacy and Information Technology -- The Years Ahead. Computers and Privacy in the Next Decade, ed. by L. J. Hoffman. New York: Academic Press.