Statistical Disclosure Limitation For Data Access

STEPHEN E. Fienberg¹, JIAshun Jin¹,²
¹Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA
²Department of Statistics, Purdue University, West Lafayette, IN, USA

Synonyms
Confidentiality protection; Multiplicity; Privacy protection; Restricted data; Risk-utility tradeoff

Definition
Statistical Disclosure Limitation refers to the broad array of methods used to protect confidentiality of statistical data, i.e., fulfilling an obligation to data providers or respondents not to transmit their information to an unauthorized party. Data Access refers to complementary obligations of statistical agencies and others to provide information for statistical purposes without violating promises of confidentiality.

Historical Background
Starting in the early twentieth century, U.S. government statistical agencies worked to develop approaches for the protection of the confidentiality of data gathered on individuals and organizations. Because such agencies also have a public obligation to use the data for the public good, they have developed both a culture of confidentiality protection and a set of statistical techniques to assure that data are released in a form that limits the identification of individual data providers [2]. In a now classic 1977 paper, Dalenius [3] described the probabilistic notion of a disclosure: “If the release of the statistics S makes it possible to determine the value [of confidential statistical data] more accurately than is possible without access to S, a disclosure has taken place.” The ensuing statistical literature on disclosure limitation has built on this probabilistic notion.

Scientific Fundamentals

Privacy, Confidentiality, and Individual Identification
Massive databases and widespread data collection and processing offer enormous opportunities for statistical analyses, advances in the understanding of social and health problems, and benefits to society more broadly. But the explosion of computerized databases containing financial and health care records and the vulnerability of databases accessible via the Internet has heightened public attention and generated fears regarding the privacy of personal data. Identify theft and sensitive data disclosure may be just a click away from a new generation of computer users and potential intruders.

Data collected directly under government auspices or at public expense are in essence a public good; legitimate analysts wish to utilize the information available in such databases for statistical purposes. Thus society’s challenge is how to release the maximal amount of information without undue risk of disclosure of individually identifiable information. Assessing this tradeoff is inherently a statistical matter as is the development of methods to limit disclosure risk. What distinguishes the field of statistical disclosure limitation from many other approaches to privacy protection is the ultimate goal of data access and enhanced data utility.

The term privacy is used both in ordinary language and in legal contexts with a multiplicity of meanings. Among these is the concept of privacy as “the right to be let alone,” e.g., see Warren and Brandeis [14], and privacy in the context of data as the control over information about oneself. But privacy is personal and subjective, varies from one person to another, and varies with time and occasion depending on the context. It is even more difficult to define precisely the meaning of “privacy-preserving” with respect to databases and the data pertaining to individual entities contained therein.
Confidentiality is the agreement, explicit or implicit, between data subject and data collector regarding the extent to which access by others to personal information is allowed. Confidentiality protection has meaning only when the data collector can deliver on its promise to the data provider or respondent. Confidentiality can be accorded to both individuals and organizations; for the individual, it is rooted in the right to privacy (i.e., the right of individuals to control the dissemination of information about themselves), whereas for establishments and organizations, there are more limited rights to protection, e.g., in connection with commercial secrets.

Disclosure relates to inappropriate attribution of information to a data provider or intruder, whether to an individual or organization. There are basically two types of disclosure, identity and attribute. An identity disclosure occurs if the data provider is identifiable from the data release. An attribute disclosure occurs when the released data make it possible to infer the characteristics of an individual data provider more accurately than would have otherwise been possible. The usual way to achieve attribute disclosure is through identity disclosure; first one identifies an individual through some combination of variables and then one associates with that individual values of other variables included in the released data.

Statistical disclosure limitation (SDL) is a set of techniques designed to “limit” the extent to which databases can be used to glean identifiable information about individuals or organizations. The dual goals of SDL are to assure that, based on released data, respondents can be identified only with relatively low probability, but also to release data that are suitable for non-identifiable analytical statistical purposes.

The Intruder

To protect the confidentiality of statistical data, one needs to understand what intruders or data snoopers want and how they may learn information about individuals in a database that require protection. Intruders may be those with legitimate access to databases and or those who gain access to a database by breaking security measures designed to keep them out. In either case, one needs to distinguish among

- Intruders with a specific target, e.g., a friend or relative. The intruder may already know that the respondent is included in the database, will possess information about the target (e.g., height, weight, habits, income) and will search the database in order to learn additional information, e.g., drug and alcohol use.
- Intruders in possession of data on multiple individuals whose goal is record linkage, e.g., to build a larger database containing more individual information. Data consolidators or aggregators fit within this category.
- Intruders without any specific target, whose goal is to embarrass the data owner. The intruder may be an enemy agent or a “hacker” eager to demonstrate a capability of breaking through efforts to limit disclosure.

Data owners can be successful in protecting the confidentiality of released data if the intruder remains sufficiently uncertain about a protected target value after data release. Various authors in the SDL literature discuss confidentiality protection from the perspective of protecting against intruders or data snoopers, e.g., see [8,12], especially those using record linkage methods [11] for attempting to identify individuals in databases.

One may consider the intruder as someone engaged in a form of a large number of statistical tests, each at significance level $\alpha$ associated with an effort to identify an individual in the database. The data owner needs to account for this somehow. Some of the null hypotheses will eventually be rejected whether or not they are actually false and thus there is a problem for the intruder as well. For the data owner, simply controlling the probability of erroneously identifying each respondent, at say 1%, is not enough. To ensure that the probability that at most one out of 1,000 individuals in a sample will be identified to be less than 1%, one in fact needs to assure that the probability of identifying each individual is no greater than $\approx 1%/1000 = 10^{-3}$.

Statistical Analysis Methods for Protecting Privacy

Matrix Masking refers to a class of SDL methods used to protect confidentiality of statistical data, transforming an $n \times p$ (cases by variables) data matrix $Z$ through pre- and post-multiplication and the possible addition of noise. The four most common forms of masking are:

1. Sampling clearly provides a measure of direct protection from disclosure provided that there is no
information of which individuals or units are included in the sample. An intruder wishing to identify an individual in the sample and link that person’s information to data in external files, using “key” variables such as age and geography available in both databases, needs to determine whether a record is unique in the sample, and then if so, the extent to which a record that is unique in the sample is also unique in the population. For continuous variables, virtually all individuals are unique in the sample, and one needs to understand the probability that an intruder would correctly match records, e.g., in the presence of error in the key variables (e.g., see Fienberg et al. [8]). For categorical data, uniqueness corresponds to counts of “1” and various authors have shown, roughly speaking, that the probability that an individual record that is unique in the sample is also unique in the population from which the sample was drawn equals the sampling fraction, $n/N$, e.g., see [7]. Thus for a sample of size 2,000 drawn from a population of 200,000,000 adults the sampling fraction is $2,000/200,000,000$ or 0.00001. The bottom line therefore is that sampling protects, just not absolutely.

2. **Perturbation** is an approach to data masking in which the transformation involves random perturbations of the original data, either through the addition of noise or via some form of restricted randomization. The simplest form of perturbation is the addition of noise. Common forms for the noise are observations drawn from a normal distribution with zero mean or perhaps a double exponential, also centered at zero. Someone analyzing the resulting transformed data must statistically reverse the noise addition process using methods from the literature on measurement error models – this requires release of the parameters of the noise component, e.g., the error variance in the normal case. Other examples of perturbation include data swapping and related tabular adjustment approaches, e.g., see [9].

3. **Collapsing** is also referred to using the labels micro-aggregation and global recoding in the statistical literature [15], and $k$-anonymity in the computer science literature. In the statistical literature on tabular categorical data, collapsing across variables in a table produces a marginal table and a popular form of data release to protect confidentiality is the release of multiple marginal tables, especially when they correspond to the minimal sufficient statistics of a log-linear model. For more details, see [10].

4. **Synthetic data** are used to replace a database by a similar one, for which the individuals are generated through some statistical process. This can be achieved through the repeated application of data swapping, e.g., see [9], or the method known as multiple imputation, e.g., [13].

Implicit in all of these techniques is the notion that when masked data are released they can be used by responsible analysts to carry out statistical analyses so that they can reach conclusions similar to those that they would have reached had they analyzed the original data. This means that all of the details of the transformation, both stochastic and non-stochastic, must be made available to the user, a point not well understood in the computer science literature or by many statistical agencies. See the related discussion in [6]. Even when one has applied a mask to a data set, the possibilities of both identity and attribute disclosure remain, although the risks may be substantially diminished. Thus one must still assess the extent of risk posed by the transformed data.

**Putting SDL Methods to Use: Risk-Utility Tradeoff**

If one is adding noise to a set of observations in order to protect confidentiality, how much noise is sufficient? And can one add too much? Clearly, too much noise will distort the data substantially and even if the details of the error variance are released the masking may impede legitimate statistical analyses of the data. The same is true for any of the methods of SDL. Thus one faces a tradeoff between data protection and data utility, something that one can assess formally using statistical decision analysis and depict graphically, e.g., see the chapter by Duncan et al. in [5]. For a slightly less formal approach to the tradeoff for categorical data protection through the release of multiple margins, see [10].

A crucial but relatively rarely discussed aspect of the risk-utility tradeoff involves the issue of multiplicity, introduced above. For illustration, consider a data set with information on 2,000 individuals, for each of which records the diagnostic result of an HIV test, with 1 corresponding to a positive result and 0 to a negative one. To protect the confidentiality for those
individuals with positive HIV test outcomes, the data owner adds noise to each record value.

Suppose an intruder wishes to identify the individuals corresponding to the proportion \( \varepsilon \) of “1”s (\( 0 < \varepsilon < 1 \)) and the data owner attempts to protect the records by adding i.i.d. Gaussian noise \( N(0, \sigma^2) \) to each data point. Clearly \( \sigma \) needs to be large enough to disguise some of the 1s and make them had to distinguish from some of the 0’s. Suppose that the intruder wants to make sure that most of those identified as “1”s are indeed “1”s (otherwise the attack on the database would be unsuccessful). In statistical terms, this means that the intruder must control for the False Discover Rate (FDR), i.e.,

\[
\frac{\text{err}}{\#(1\text{-class})} = \text{FDR}.
\]

Consider an intruder who decides to set the FDR at 5% by picking a threshold \( \sigma_t \) and classifying any entry as a “1” if the observed value exceeds the threshold. By elementary statistics, the number of misclassified “1”s is distributed as a binomial random variable, \( B(n(1 - \varepsilon), \Phi(t)) \), and the number of correctly classified “1”s is distributed as \( B(n(1 - \varepsilon), \Phi(1 - \frac{1}{\varepsilon})) \). Consequently, the associated FDR is

\[
\text{FDR} \approx \frac{n(1-\varepsilon)\Phi(t)}{n(1-\varepsilon)\Phi(t) + n\Phi(1-\frac{1}{\varepsilon})} = 1 + \left(\frac{1}{\Phi(1-\frac{1}{\varepsilon})}\right)^{\frac{1}{n}} - 1,
\]

where \( \Phi = 1 - \Phi \) is the survival function of the standard normal distribution function.

Consider a high risk population where \( 50\% \) of the individuals test positive for HIV, i.e., \( \varepsilon = 0.05 \), and the data owner uses a similar level of noise addition for confidentiality protection, i.e., \( \sigma = 1 \). Then to ensure that FDR \( \leq 5\% \), the intruder needs to set \( \Phi(1-\frac{1}{\varepsilon}) = 0.95 \), which yields \( t \approx 2.49 \). Setting the threshold high enough so that the chance for each of the individuals exhibiting a “1” to be correctly classified as “1” is \( \Phi(t-1) = \Phi(2.32) \approx 0.0165 \), which seems not very large. But since \( n = 2000 \), the number of “0”s that are misclassified as “1”s is approximately \( n(1-\varepsilon)\Phi(3.132) = 2000 \times 0.0165 = 33 \), and the number of “1”s that are correctly classified as “1”s is approximately \( n\Phi(2.132) = 2000 \times 0.0165 \approx 33 \). This says that the intruder is able to identify 17 records, out of which 16 are corrected classified!

Alternatively, one might ask about the probability that no more than \( k \) “1”s are correctly classified, i.e.,

\[
\sum_{i=0}^{k} \binom{n}{i} p^i (1 - p)^{n-i}, \quad p = \Phi(2.132).
\]

For \( k = 0, 3, 6, 9 \), the probabilities are correspondingly \( 5.45 \times 10^{-8}, 5.22 \times 10^{-5}, 2.62 \times 10^{-3}, 0.031 \). To understand the implications of these values, consider \( k = 9 \). This says that with probability as high as 97\% there are actually “1”s that are correctly identified as “1”! For many this might seem to be a worrisome situation, and it raises issues associated with the efficiency of adding noise that have not appeared in the statistical literature on confidentiality protection.

Suppose that the data come from a low risk population where only 5\% or 10\% individuals test positive for HIV, i.e., \( \varepsilon = 0.05 \), and the data owner uses a similar level of noise addition for confidentiality protection, i.e., \( \sigma = 1 \). Then to ensure that FDR \( \leq 5\% \), the intruder needs to set \( \Phi(1-\frac{1}{\varepsilon}) = 0.95 \), which yields \( t \approx 6.22 \). Correspondingly, \( \Phi(t) = 2.5 \times 10^{-10} \) and \( \Phi(t - \frac{1}{\varepsilon}) = 8.9 \times 10^{-8} \). The expected number of “0” that are misclassified as “1”s is approximately \( n(1 - \varepsilon)\Phi(6.22) = 2000 \times 0.05 \times 8.9 \times 10^{-8} \approx 9.5 \times 10^{-6} \). In this case, since \( n\Phi(6.22) \ll 1 \), the approximation is inaccurate and one needs to take a different approach.

In fact, the proportion of true HIV cases is so small that the example falls into the so-called very sparse regime studied in detail in the multiple testing literature, see for example [1,4]. One phenomena from that literature implies that when the noise level is relatively high, the extreme values are not necessary related to cases with positive HIV tests. Consider the following simulated data set with \( n = 2000 \) cases, where 100 of them are HIV (equal to 1) and all others are non-HIV (equal to 0). The data owner adds independent standard Gaussian \( N(0,1) \) noise to each value. Figure 1 shows the result where red correspond to cases with positive HIV tests, and green correspond to cases with negative HIV tests. The red values are larger than typical green ones, but not larger than all of them. In fact, among the largest 10 values, only 2 are red, with 8 are green.

This leads us to another interesting phenomenon from the statistical literature on the FDR. Let \( m_{\text{FDR}} \) denote the minimum FDR across all possible thresholds \( t \), \( m_{\text{FDR}} = \min_{t \in [0,1]} \{ \text{FDR}_t \} \). How small can \( m_{\text{FDR}} \) be? Figure 1 shows the histogram of the \( m_{\text{FDR}} \) values for 100 independent repetitions of the simulation experiment. More than half of the time, the \( m_{\text{FDR}} \) value is no less than 15\%, and sometimes it is as great as 50\% and larger!
Figure 1. Left Panel: Perturbed HIV data through addition of independent draws from $N(0,1)$. Those values associated with positive HIV tests are in red, and those with negative HIV tests are in green. Right Panel: 100 simulated mFDR values based on 100 simulation for $n = 2,000$ and $\varepsilon = .05$ and added noise from $N(0,1)$. 
This simple example implies that with $\sigma = 1$, the noise level might be so large that the intruder cannot correctly identify any HIV cases. But from the perspective of the risk-utility tradeoff, one also needs to ask whether the noise level is so high that the data are no longer analytically useful. Thus one needs to ask: What is the largest noise variance that still allows for valid inferences, c.f., [1,4]. If the number of true HIV cases is

$$m = m_n = n^{1-\beta}, \quad (1)$$

and the noise level is $\sigma = \sigma_n = \frac{1}{\sqrt{2r\log n}}$, where $0 < \beta$, $r < 1$ are parameters, then as $n$ tends to $\infty$, there is a boundary, $r = \beta$, which separates the $\beta$-$r$ plane into two regions: the classifiable region and the non-classifiable region; In the interior of the classifiable region, asymptotically, it is possible to isolate completely the cases with positive HIV tests from those with negative ones. In fact, there is a threshold by which one can identify that subset of the data corresponding to positive HIV tests. On the one hand, almost every “identified” HIV case has a positive HIV test and the subset includes almost all the cases with positive HIV tests. In the interior of the non-classifiable region, by contrast, such isolation of cases is impossible. In fact, given any chosen threshold, either one situation or the other occurs!

For the example of $n = 2,000$ and $m = 100$ cases with positive HIV tests. take $\beta = 1 - \log(m)/\log(n) \approx 0.3941$ in model (1). In order not to have complete isolation of cases with HIV, one should take $\sigma_n > \frac{1}{\sqrt{2\beta\log(n)}} \approx 0.409$. For $\sigma_n = 0.5$, consider a repetition of the case of $n = 2,000$ and $\varepsilon = 1/2$, as well as the case of $n = 2,000$ and $\varepsilon = 0.05$. Thus $1/\sigma = 2$ and to control the FDR at $5\%$, one must evaluate

$$\frac{1}{\sqrt{2\beta\log(n)}} = 19,$$

which yields $t = 0.54$. Correspondingly, $n(1 - \varepsilon)\Phi(t) = 2,000 \times 0.54 \approx 29$, and $n(1 - \varepsilon)\Phi(t - 2) = 2,000 \times 0.25 \approx 50$. Furthermore, for $k = 0,3,6,9$, the probability that no more than $k$ “1”s are correctly classified are extremely small ($< 10^{-31}$). For the case of $n = 2,000$ and $\varepsilon = 0.05$. Similarly, one must similarly evaluate

$$\frac{1}{\sqrt{2\beta\log(n)}} = 3.62,$$

which yields $t = 3.62$. Correspondingly, $n(1 - \varepsilon)\Phi(t) = 2,000 \times 0.95 \times 1.47 \times 10^{-4} \approx 0.28$, and $n(1 - \varepsilon)\Phi(t - 2) = 2,000 \times 0.05 \times 0.053 \approx 5$. Furthermore, for $k = 0,3,6,9$, the probability that no more than $k$ “1”s are correctly classified are $4.5 \times 10^{-3}$, $0.22, 0.73, 0.96$. Take $k = 3$. The probability that more than three “1”s are correctly classified is about 78%.

This example illustrates how the multiplicity issue arises when an intruder tries to match many records with those in a database “protected” by matrix masking. Simply protecting each individual record with high probability is not enough; there remains a substantial chance for one or more record to be vulnerable to disclosure.

**Summary**

The accumulation of massive data sets and the rapid development of the Internet expanded opportunities for data analyses as well as created enormous challenges for privacy protection. This brief overview of the literature on statistical disclosure limitation has stressed four categories of approaches: sampling, perturbation, collapsing (aggregation), and the use of synthetic data. An overlooked issue in privacy protection is the notion of “multiplicity,” which is present whenever one attempts to protect many records simultaneously, or a data intruder tries to match the records of multiple targets in a database simultaneously. Simply protecting each individual record with high probability does not automatically protect all records, and careful statistical measures for resolving the multiplicity issue are necessary.

**Key Applications**

The methods of statistical disclosure limitation outlined here are already in widespread use by government statistical agencies throughout the world. The volumes by Doyle et al. [5] and Willenborg and de Waal [15] summarize a number of the approaches and methodologies. In particular, census data released by most developed countries are protected using these methods.

**Future Directions**

The elaboration of approaches described here to deal with very large scale databases remains a challenge, especially in the face of demands for increased access to data and novel attacks on databases by intruders. This entry presents the first known application of ideas and results from the multiplicity literature to the problem of statistical disclosure limitation and risk-utility tradeoff. These ideas further development and integration with the rest of the literature. And methodology will

**Cross-references**

- Data Perturbation
- Individually Identifiable Data
Inference Control in Statistical Databases
Matrix Masking
Privacy
Privacy-Preserving Data Mining
Randomization Methods to Ensure Data Privacy

Recommended Reading