Statistics Surveys
Vol. 5 (2011) 1–29
ISSN: 1935-7516
DOI: 10.1214/11-SS074

Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy*

Gregory J. Matthews and Ofer Harel†

Department of Statistics, University of Connecticut
215 Glenbrook Rd. U-4120, Storrs, CT 06269
e-mail: gjm112@gmail.com; oharel@stat.uconn.edu

Abstract: There is an ever increasing demand from researchers for access to useful microdata files. However, there are also growing concerns regarding the privacy of the individuals contained in the microdata. Ideally, microdata could be released in such a way that a balance between usefulness of the data and privacy is struck. This paper presents a review of proposed methods of statistical disclosure control and techniques for assessing the privacy of such methods under different definitions of disclosure.

AMS 2000 subject classifications: Primary 62A01.
Keywords and phrases: Confidentiality, Privacy, Disclosure Limitation, Missing Data, Synthetic Data, Multiple Imputation, Differential Privacy.

Received May 2010.

1. Introduction

Article 12 of the Universal Declaration of Human Rights (General Assembly of the United Nations, 1948) states: "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks." As such, with privacy being viewed as a basic human right by the United Nations, data releasing agencies must make every effort possible to maintain high levels of privacy for the individuals who entrust their data to an agency.

What exactly is meant by privacy? Given a piece of information about an individual, one person may wish to keep that data private while another individual may not particularly care about that specific piece of information. This leads to a good definition of privacy. Fellegi (1972, page 7) used the definition of privacy provided by Professor Weston of Columbia University which defines privacy as the right "to determine what information about ourselves we will share with others."

*This paper was accepted by Louis-Paul Rivest, Associate Editor for SSC.
†Corresponding author

Privacy considerations of microdata are an increasingly important issue. The amount of data being produced everyday pertaining to individuals is unprecedented. Between medical, educational, and human services records, large amounts of data are produced. These types of data are invaluable to researchers in a vast array of fields, driving demand for this data. However, this raw data cannot simply be released to the public for study due to these privacy concerns.

Many agencies rely on publicly released data from the census, and numerous public policy research projects depend on publicly available medical or educational data sets. Further, agencies like the U.S. National Institute of Health (NIH) urge its data collecting grantees to release their data for public use, but they require that this be done in a private way. They state: "In NIH's view, all data should be considered for data sharing. Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data. To facilitate data sharing, investigators submitting a research application requesting $500,000 or more of direct costs in any single year to NIH on or after October 1, 2003 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible."

Often times, the most interesting data for research can be extremely sensitive information about an individual that must remain private for ethical or even legal reasons (e.g. Health Insurance Portability and Accountability Act (HIPAA), Family Educational Rights and Privacy Act (FERPA)). HIPAA creates a legal protection for individuals who wish to keep their medical records private, whereas, FERPA provides individuals with legal protection of their educational data. Data collecting organizations have a further incentive to maintain the privacy of their respondents' data that goes beyond ethics or the law: If respondents feel that their data are at risk for disclosure, they may be less likely to be completely honest in their responses. This may cause respondents to alter responses or simply not respond at all to some surveys. Therefore, trust between a data collecting agency and its respondents is very important.

Ideally, any useful collected data set could be released to the public for research with the implicit trust that that the data would not be used for inappropriate purposes. However, groups or individuals often have incentives to use data maliciously. For example, in 1995, prior to the passage of HIPAA, Woodward (1995) described a case involving a banker from Maryland who obtained a list of patients with cancer. Using the list of patients with cancer along with a list of clients with outstanding loans, the banker sought to match individuals across both lists. When a match was found, he then called in the loans of the clients who had cancer. Today, with the regulations of HIPAA, private medical information cannot simply be released to the public. As such, institutions that wish to release sensitive data must take steps to protect the identity of the individuals in the data.

The first, most basic step in maintaining privacy is to remove variables such as name, social security number, and home address. Agencies strive to do their best to de-identify the data so that the privacy of the individual remains in-

tact, while still providing researchers with useful data with which they can use to make useful, correct conclusions. However, simply removing these obvious identifiers is not always enough to maintain the privacy of an individual. For instance, several years ago the Massachusetts Group Insurance Commission released data to the public for research that was stripped of obvious identifers. Sweeney (2002b) used this data, along with publicly available voting records, to identify the released medical information or former Massachusetts governor William Weld.

Sweeney (2002b, page 2) went on to say "...87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. Clearly, data released containing such information about these individuals should not be considered anonymous. Yet, health and other person-specific data are often publicly available in this form." Thus, simply removing obvious identifiers from the data is not always adequate to maintain the privacy of the individual. More rigorous procedures are required to achieve privacy.

It is this type of disclosure, from what Sarathy and Muralidhar (2002a) referred to as "snoopers", that is discussed here. (As opposed to, say, privacy breaches from unauthorized users of a database (hackers).) Sarathy and Muralidhar (2002a, page 1) stated: "The security threat posed by snoopers generally takes the form of undesired inferences about confidential data using other data available either within or outside the database."

We view all data discussed as rectangular data with each row representing an observation and each column representing a variable, however, the rectangle need not be complete. For some methods, rectangular data is expressed in tabular format, and the discussed techniques for tabular data would be applied.

While we consider this to be a thorough review, the breadth of the topic is vast, and we do not attempt to cover all papers on the topic. Another very good review of disclosure control techniques which protect against this type of disclosure can be found in Skinner (2009).

This manuscript discusses methods for limiting statistical disclosure in Section 2. Section 3 discusses measures for assessing the privacy of statistical disclosure control techniques, while Section 4 concludes the manuscript with a summary.

2. Releasing microdata to the public in a private way

Microdata are data containing observations on individual level. When this type of data is released for research purposes the very first action taken to maintain confidentiality is the removal of obvious identifiers such as name, address, social security number, zip code, etc. However, as mentioned above, this is not always enough to protect the privacy of the individual from an inferential disclosure which can occur, for example, when an individual in the released microdata has some outlying or unique trait (e.g. a very large income, a rare occupation).

In this section, we discuss different proposed privacy preserving techniques for releasing data for research. We start by discussing basic privacy preserving methods employed by agencies for releasing data. This is followed by several other proposals for maintaining privacy, including matrix masking, data swapping, and synthetic data. Adam and Worthmann (1989) and Duncan and Pearson (1991) both presented good reviews of some of the methods mentioned in this section.

2.1. Basic methods for limiting disclosure risk

After removing obvious identifiers, some of the most basic methods for maintaining privacy of publicly released data sets employed by data releasing agencies (e.g. The U.S. Census Bureau) include limitation of detail, top/bottom coding, cell suppression, and rounding.

1. Limitation of detail: This technique includes recoding variables into intervals and collapsing together categories in which only a small number of observations appear. For example, the U. S. Census does not release geographic identifiers that would leave a sub-population with less that 100,000 observations (Moore, 1996)
2. Top/bottom coding: This technique can help reduce the disclosure risk of extreme values in the data by limiting the largest (or smallest) value possible for a given variable. For example, if an individual has an extremely large salary, rather than reporting the exact amount, which would make the observation vulnerable to disclosure, an agency may simply report it as "over $100,000". Likewise, negative values of income could be recoded to be "less than $0" to avoid extremely large negative values.
3. Suppression: In a contingency table, cells with too few observations cannot be released to the public, as it may be easy to to infer the identity of these individuals. A simple procedure for controlling disclosure is suppression of these cells. Similarly, if the values of some combination of variables are unique or nearly unique in the data, the identity of these rare combination may be easily de-identified. Therefore, these observations could be suppressed as one possible method for maintaining confidentiality. (Cox, 1980, 1984, Mugge, 1983, Cox et al., 1987)
4. Rounding: Rounding is another method to limit statistical disclosure of data. Random rounding involves deciding on a rounding base and then rounding each observation up or down to the nearest multiple of the rounding base. Rounding up or down is decided upon randomly based on how close the observation is to the nearest multiple of a rounding base. For example, if the rounding base is 10 and 7 was observed, 7 would be rounded up with probability 0.7 and rounded down with probability 0.3. One could also use controlled rounding which allows the sum of the rounded values to be the same as the rounded value of the sum of the original data. (Cox, 1984, Cox et al., 1987, Cox, 1987)

5. Addition of noise: Rather than release the actual values of the data, noise is added to the data in an attempt to prevent a linkage attack from occurring. The perturbed data can be correctly analyzed by accounting for the extra variability from the added noise. For continuous data, noise addition is discussed in Fuller (1993) and for discrete data a technique called the Post Randomization Method (PRAM) Gouweleeuw et al. (1998) can be applied.

2.2. Sampling

Sampling is a very powerful tool in limiting disclosure risk of released microdata files, especially against linkage attacks. For instance, a malicious user may try to match an observation in a released set of microdata to another observation in a data set which could identify the individual. However, simply by matching a record in the released data file does not mean that the match is correct. Skinner et al. (1994) pointed out that "Population uniqueness will be a sufficient condition for an exact match to be verified as correct." If the released microdata are a sample, this make it difficult to verify population uniqueness and is one of the key benefits of sampling.

Other benefits of sampling as method of disclosure control are that it is easy to implement and the resulting sampled data are relatively easy to analyze.

2.3. Matrix masking

Cox (1980) and Cox (1994) proposed a statistical disclosure limitation (SDL) method called matrix masking. Consider an n by p data matrix, X, consisting of n observation and p variables. Rather than release the data X, one could release the data $Y = AXB + C$ where A, B, C are appropriate conformable matrices. By properly defining the matrices A, B, and C, special cases of matrix masking include: noise addition (Fuller, 1993), sampling, suppressing sensitive variables, cell suppression, and addition of simulated data.

A drawback to matrix masking is that in order to analyze the data, the analyzer must have knowledge of the masking procedure used, and, often, even if the consumer knows the masking procedure, the analysis of the data can be complex and special software may be needed. Analysis of masked data is discussed in Little (1993).

Kim (1986) proposed to protect microdata via the addition of noise and transformation. Using their notation, for a data set, x, consisting of n observations and p variables. Kim (1986) suggested masking the j-th variable, x_j by adding noise, e_j, from a normal distribution or from the distribution of x_j itself. Thus the masked, released data for the i-th observation of the j-th variable, y_{ij} will be $x_{ij} + e_{ij}$ where $i = 1...n$ and $j = 1...p$. Kim (1986) further suggests a transformation after the addition of noise of the form $z_{ij} = ay_{ij} + b_j$ where a and b_j are chosen subject to constrains on the first and second moments of z_j and y_j. b_j is chosen such that $E[x_j] = E[z_j]$ and a can either be chosen so that

$Var[x_j] = Var[z_j]$ or based on the specific confidentiality requirements of the application.

While Kim (1986) discussed many of the properties of this masking procedure, it does not explore the degree to which disclosure is limited leaving this as a topic for future work. It notes, however, that by properly controlling the value of a in the transformation the probability of re-identification can be raised or lowered as appropriate. Further, when using this method the bivariate relationships remain intact and common analyses, like regression, for example, using the transformed data will perform well.

Bowden and Sim (1992) introduced what they refer to as the privacy bootstrap. Rather than adding random noise from a known distribution, the added noise is based on the empirical distribution of the data via a bootstrapping procedure. Consider the actual data to be $x_i, i = 1, ..., N$ with mean $\bar{x}$ and let $x_i^\star$ be a randomly sampled observation from the collection of $x_i, i = 1, ..., N$ where each observation has probability $\frac{1}{N}$ of being selected. Then the released data would be $X_i = x_i + \epsilon_i^\star$ where $\epsilon^\star = x_i^\star - \bar{x}$. One could also choose to release data as $X_i = x_i + \alpha\epsilon_i^\star$ or $X_i = \beta x_i + \alpha\epsilon_i^\star$ with α and β chosen based on the specific situation.

2.3.1. Randomized response and Post Randomization Method (PRAM)

Randomized response (Warner, 1965, Greenberg et al., 1969) is a technique used in surveys when the questions being posed are of a sensitive nature (Suppose an interviewer was asking about illegal activity which, in turn, may make the respondent more likely to lie or simply refuse to respond.). The basic idea is that a respondent answers a question truthfully with some probability p or answers the question untruthfully with probability $1 - p$. In this way, the survey taker does not know for sure whether the respondent is telling the truth or not and a level of confidentiality is maintained. Surveys with randomized response were originally proposed to remove the effect of response bias in surveys that ask sensitive questions.

By using this technique respondents privacy is protected, since, even if an individual is identified by a data snooper, they cannot be sure whether the response is correct or not. For example, when administering a survey a researcher may ask a question which would easily identify the respondent, such as asking about a rare condition or disease. After the question is asked, the respondent flips a coin and, for example, tells the truth when heads is observed and lies when tails in observed. In this way, even the raw microdata maintains a level of confidentiality. This method could also be applied after raw microdata were collected. For each observation, the real value of a sensitive field would be released with some probability and its opposite would be released with some other probability. Either way, in order to analyze this data, the researcher must have information about the randomization mechanism.

Gouweleeuw et al. (1998) introduced Post Randomization Method which is used to protect categorical data from disclosure. PRAM perturbs each record in

a data file using some probability distribution. This essentially amounts to the addition of noise for categorical variables. One important distinction between PRAM and randomized response is that in randomized response the random mechanism is independent of the true score and applied at the time of collection. However, with PRAM the true value is known and one can therefore condition on this value when defining the probability mechanism used to perturb the data.

2.4. Data swapping and data shuffling

Data swapping was first proposed by Dalenius and Reiss (1982) as a method of disclosure limitation. The proposed procedure was intended to be used for contingency tables within a database. Then, as the name implies, the data are swapped in such a way as to maintain the marginal counts of the table. The swapping procedure adds a layer of protection, while the marginal counts remain intact. Dalenius and Denning (1982) also suggest the possibility of releasing the moments of continuous data rather than the data itself.

Moore (1996) identified several desirable properties of data swapping. First, the procedure allows information about each respondent to be masked. Also, swapping only needs to be performed on sensitive variables in order to remove the relationship between the record and the respondent. This leaves non-sensitive variables undisturbed. Finally, as a practical consideration, Moore (1996) noted that the procedure is easy to implement, requiring only a microdata file and a random number generator.

If one is simply interested in univariate statistics, this procedure works very well, however, one drawback to the procedure is that it may not maintain multivariate relationships. Also, it is likely that analysis of sub-populations may be affected by the swapping procedure. It is also possible that the swapping may result in nonsensical combinations. For example, if your data contains gender and type of cancer, after a swap, the resultant data may contain a record indicating there is a female with prostate cancer.

Tables 1A and 1B offer an example of the implementation of data swapping from Fienberg and McIntyre (2004). Table 1A contains the original unperturbed microdata, while Table 1B displays the data after data swapping has occurred. Here, X is the sensitive variable, so data swapping is only performed on X, while Y and Z remain the same in the original and swapped data.

While the original intention of data swapping was to be used for releasing contingency tables of the swapped data, the problem can be extended for microdata. However, if one wishes to release microdata many more swaps must be made to preserve the level of privacy. Identifying the correct number of swaps, Fienberg and McIntyre (2004) noted, is "computationally impractical." As such, it is suggested that the counts be preserved only approximately. This idea is discussed in detail in Reiss (1984). Also, data swapping makes it very difficult to maintain weighted counts when the weights are unequal, which occurs often in surveys.

TABLE 1
An example of data swapping. (A) contains the unswapped, original values of the data. (B) presents the data after data swapping.

(A) Raw Data				(B) Swapped Data			
Record	X	Y	Z	Record	X	Y	Z
1	0	1	0	1	1	1	0
2	0	1	0	2	0	1	0
3	0	0	1	3	0	0	1
4	0	0	1	4	1	0	1
5	1	1	1	5	0	1	1
6	1	0	0	6	1	0	0
7	1	0	0	7	0	0	0

Liew et al. (1985) proposed a swapping method where the released data are random draws and not the original variables. This method requires the identification of the univariate distribution of each variable which is considered to be sensitive for release to the public.

Carlson and Salabasis (2002) proposed a procedure that they refer to as the C&S method which offers an improvement to the proposal put forth in Liew et al. (1985). While they show that their swapping method maintains a large amount of utility, they make no claims or observations as to the confidentiality of their method. Later, Sarathy and Muralidhar (2002a) showed that the C&S method has almost no desirable properties as a method for limiting statistical disclosure.

Moore (1996) outlined a method called rank based proximity swapping which was proposed in an unpublished article in Greenberg (1987). This procedure can be used for masking data as long as the variables of interest are continuous in nature. The main difference between the Greenberg (1987) swapping procedure and Dalenius and Reiss (1982) proposal is that the range over which the data can be swapped is restricted. The advantage here is that by limiting what values can be swapped with other values, many of the multivariate relationships can be more appropriately maintained, whereas with Dalenius and Reiss (1982) swapping, these relationships may be lost.

Sarathy and Muralidhar (2002a) went on to propose a further method called data shuffling based on the conditional distribution approach. In proposing this, they seek a method that performs as well as data swapping, but without the inherent disclosure risks. Under their method, as in data swapping, all of the marginal distributions remain intact. They also show that pairwise monotonic relationships in the original data are maintained in the released shuffled data. They also note that releasing shuffled data does not increase the possibility of disclosure even when shuffled microdata are released.

2.5. Synthetic data

Synthetic data, first proposed by Rubin (1993), is a method of statistical disclosure limitation based on the missing data technique multiple imputation (Rubin,

1987, Little and Rubin, 1987, Schafer and Graham, 2002, Harel and Zhou, 2007). The idea is to view sensitive data as missing values and replace them using multiple imputation techniques. Thus sensitive attributes would be replaced by random draws from an appropriate posterior predictive distribution.

One can think of the observed microdata as a random sample of size n from a population P of size N. The population is made up of background variables of interest, $X = (X_i, i = 1, 2, ..., N)$, which might include name, birthdate, address, etc. and survey variables of interest, $Y = (Y_i, i = 1, 2, ..., N)$. We randomly take a sample of size n from the population which yields $Y_{inc} = (Y_{inc,i}, i = 1, 2, ..., n)$. Therefore, the observed microdata D consists of the background variables, X, and the observed survey variables, Y_{inc}. The remaining $N - n$ unsurveyed individuals make up $Y_{exc} = (Y_{exc,j}, j = n + 1, n + 2, ..., N)$. Next, multiple imputation is used to replace Y_{exc} with plausible values. These imputations are drawn from the posterior predictive distribution $Pr(Y_{exc}|Y_{inc}, X)$. This process is repeated M times, each time creating a synthetic population $P^{(l)}$ of size N with $l = 1, 2, ..., M$. A random sample of size k is then drawn from each synthetic population, $P^{(l)}$, yielding $D^{(l)}$ with $l = 1, 2, ..., M$. Thus the released fully synthetic data are $D_{syn} = (D^{(l)}, l = 1, 2, ..., M)$.

The data releasing agency may want to take privacy a step further and release neither X nor Y_{inc}. In this case, Raghunathan et al. (2003) recommended creating a "future" population by randomly drawing from a posterior predictive distribution of $Pr(X_f, Y_f|X, Y_{inc})$ where X_f, Y_f are random variables representing a "future" population. In this case, no actual data are released, which makes linkage attacks very difficult. Fully synthetic data sets are discussed in Raghunathan et al. (2003), Rubin (1993), Reiter (2002, 2004a,b, 2005b) and Matthews et al. (2010b).

Alternatively, an agency could employ partially synthetic data techniques as proposed in Little (1993). Rather than replacing all of the data with imputations, as is the case in the fully synthetic framework, only sensitive attributes are replaced with imputations. This can be done in many ways, including determining that an entire variable or variables must remain private or by selecting individual attributes that are at high risk of disclosure. Once an agency decides what values must remain private, they consider those values to be missing and replace them using multiple imputations techniques. This creates M partially synthetic data sets which will be released. Each partially synthetic data set consists of the non-sensitive data, which will be the same across all M synthetic data sets, and the imputed values of the sensitive data. Partially synthetic data methods are discussed in Kennickell (1997), Abowd and Woodcock (2001), Liu and Little (2002), and Reiter (2003, 2005c).

One big advantage of synthetic data is the ease with which the data can be analyzed. In classic multiple imputation, each imputed data set is analyzed using a complete data technique and the inferences are combined using the appropriate combining rules (Rubin, 1987). Analysis of synthetic data sets is performed in a similar fashion. Each synthetic data set is analyzed using a complete data technique and inferences are combined using the appropriate combining rules, which are slightly different than that of classic imputation. The combining rules

for fully and partially synthetic data are set forth in Raghunathan et al. (2003) and Reiter (2003), respectively.

A hurdle that must be overcome in dealing with synthetic data is convincing researchers that analyzing data that is not "real" has merit. Evidence demonstrating the usefulness of synthetic data is presented in Raghunathan et al. (2003) and Reiter (2005b) where they show that if the imputation model specified is accurate, many resulting analyses based on the synthetic data set will be virtually identical. However, if the model for imputation is incorrect or inaccurate, the resulting analysis from the synthetic data will yield parameter estimates that are much different than those estimated from the actual data (Reiter, 2005b, Matthews et al., 2010b). As such, synthetic data sets are only as good as the models used for imputation.

2.6. Other selected privacy preserving methods

2.6.1. Slicing, micro-aggregates, and recombination

Paass (1988) suggested slicing, micro-aggregates, and recombination as methods of controlling statistical disclosures. The first method involves taking a set of complete records and slicing them into groups, each of which would have a smaller number of variables. Then each slice is released separately. Micro-aggregation involves creating new records by averaging at least three original records. The third proposal is what they refer to as recombination. This involves dividing each record into sub-records consisting of several variables each. Then the sub-records are recombined across different individuals to create synthetic records. In order to retain the original relationships between the variables, recombinations are done in such a way that that "...only those subrecords whose underlying complete records were of similar structure were recombined (Paass, 1988, Page 493)."

Slicing separates the variables into smaller groups, rather than releasing all of them at once. This method maintains suitable levels of confidentiality, however, lacks utility for more complicated analyses. Micro-aggregation, creating new records by averaging at least three records, was found to be unsuitable, as the resulting utility of the data was substantially diminished by this disclosure avoidance technique.

The use of recombination is shown to be the best of the three methods evaluated here. This method consists of decomposing each observation into 8 sub-records of between 5 and 15 variables then the sub-records are combined with other sub-records using statistical match until all of the sub-records have been recombined. They show that this method is safe against usual disclosure attempts using additional information and that many common analyses provide suitable results.

2.6.2. Location data

Krumm (2007) looked at inference attacks on location data collected from commuters global positioning system (GPS). They showed that they can make rea-

sonable inferences as to where the commuter lived, based solely on the data, creating a possible breach of privacy. They offer several possible methods for increasing privacy, including: spatial cloaking, addition of noise, and rounding. Spatial cloaking involves suppressing all of the points within a circle surrounding the house of a commuter. However, the center of the circle is randomly chosen within some bounds near the house because simply centering the circle exactly at the location of the house would make it very easy for an intruder to elicit the exact location of interest. The addition of noise involves adding 2-dimensional noise to each point obscuring the exact location of the commuter. The third method involves rounding each point the nearest point of a grid. The coarseness of the grid can be adjusted to add more or less privacy. Armstrong et al. (1999) also discussed protecting geographic data released to the public.

2.6.3. Scrub system, Datafly, Argus, and SUDA2

Sweeney (1996) proposed the Scrub system. This algorithm scans through personal medical records to locate information which could be used to identify the owner of the records. Words or phrases which would put the owner of the record at risk are identified and replaced with a "pseudo-value". However, even after locating and replacing these identifying words and replacing them, anonymity still cannot be guaranteed. Sweeney (1996, page 5) noted "Even then however, we still cannot scrub implicit information where an overall sequence of events whose preponderance of details identify a particular individual. This is often the case in mental health data and discharge notes".

Argus (De Waal et al., 1995, Hundepool et al., Feb. 2005) is a software package for limiting the risk of statistical disclosure. The goal of Argus is to limit the occurrence of rare combinations of identifying variables, thus lowering the risk of disclosure. This is achieved using global recoding and local suppression. Global recoding involves combining several categorical variables into one. Therefore, for instance, rather than releasing the city or town that someone lives in, individuals within the data could be grouped into county or even state. This makes it more difficult to identify an individual in the data. Following global recoding, suppression is used to remove combinations of identifying variables that still appear in rare combinations.

Sweeney (1997) proposed the Datafly system. This system processes specific queries made to a database. Then the query results are returned, subject to a specified level of privacy between 0 and 1. A specified level of 0 would return the raw data from the query, and at level 1, the data would be generalized as much as possible. Privacy is achieved in two ways: 1) Data are returned in bins rather than in raw form and 2.) data which fit into a bin with too few observations is simply not returned in the query results. These two steps are essentially global recoding and local suppression as specified by Argus.

The algorithm SUDA2 (Special Unique Detection Algorithm) (Manning et al., 2008) is another useful piece of software for statistical disclosure control. This algorithm searches a data set for unique observations. One benefit of SUDA2

that the authors note is that SUDA2 allows "significantly more columns to be addressed." This is important as the the number of potential variables in data sets can be quite large.

2.7. Micro-agglomeration, Substitution, Subsampling, and Calibration (MASSC)

Micro-agglomeration, Substitution, Subsampling, and Calibration (MASSC) (Singh et al., 2003) is a combination of several individual statistical disclosure techniques. The procedure proceeds in four basic steps: micro-agglomeration, substitution, subsampling, and calibration. Micro-agglomeration refers to placing records into groups depending on the level of assessed risk. Identifying variables are broken into two categories, core and non-core variables. Core identifying variables are variables that will be easily available to an intruder, while non-core identifying variables will be less readily available. Records at the highest risk level are records that are unique in terms of core variables, while the lowest level of risk includes records which are not unique in terms of both types of variables, core and non-core. Once records have been grouped into risk categories, disclosure control techniques are applied. First, substitution techniques are used to perturb the data. Substitution refers to many disclosure control techniques including recoding, random rounding, addition of random noise, data swapping, and imputation (synthetic data). Following this step, a subsampling step is applied to add further protection to the data. Finally, the released data are calibrated such that specific estimates based on the released data match the estimates based on the original data. The authors note that this "helps reduce the bias caused by substitution" (Singh et al., 2003, Page 9). One very desirable property of MASSC is that both disclosure risk and information loss can be controlled for simultaneously.

3. Assessing privacy

In order to assess privacy, disclosure must be defined since different definitions of disclosure will lead to different definitions of privacy. Willenborg and de Waal (2001) categorized disclosure risk into two main categories, namely, the risk of re-identification and the risk of predictive disclosures. A re-identification occurs when one is able to accurately identify an individual in the released data, whereas a predictive disclosure occurs when the value of some unknown sensitive attribute can be estimated with reasonable accuracy.

Duncan and Lambert (1989) defined four types of privacy, identity disclosure, attribute disclosure, inferential disclosure, and population disclosure. Identity disclosure is, as before, being able to accurately identify and individual in the released microdata. Attribute disclosure occurs when an intruder is able to obtain "reliable information about an individual as the result of linking". Inferential disclosure occurs when a consumer of the released data is able to infer new information about an individual even without linking to a specific observation.

Finally, population, or model, disclosure occurs if a confidential information about a population can be inferred through the construction of a model based on the released microdata.

Regardless of how privacy is defined, any organization which plans on implementing disclosure control techniques in order to privately release data to the public needs to be aware of the trade-off between privacy and data utility. It is always possible to increase the privacy of any specific data release, but this almost assuredly comes with a loss of data utility. Therefore, privacy cannot be assessed by itself, it must always be measured in conjunction with the utility of the data after privacy preserving techniques have been applied.

In this section, we will first discuss measures of privacy based on the threat of re-identification and attribute disclosure. This is followed by a review of procedures for assessing privacy based on inferential disclosures and population disclosure.

3.1. Re-identification measures

Spruill (1982) discussed the confidentiality of several methods of protecting public release microdata files. First, a subset of the data was chosen and then a masking procedure was applied. The masking procedures they tested included: addition of normal random error, grouping, random rounding, and data swapping. The proposed measure of confidentiality is based around how many records in the released data can be linked to their respective data in the unmasked data. Calculation of the measure of confidentiality is as follows: An element of the data is selected and masked. The masked element is then compared to all elements in the unmasked data. The element in the unmasked data which minimizes the absolute deviations or squared error is selected. If the selected element from the unmasked data is the same as the element that generated the chosen masked element, then they say a link has been made. The measure of confidentiality is simply the percent of the elements for which a match cannot be made. Spruill (1983) presents a demonstration of the privacy measure using real data.

Paass (1988) assessed privacy based on the number of matches that can be made between some additional information and the released data. A match occurs, for their purposes, when a record in the additional information matches or nearly matches a record in the released data. Along with this, Paass (1988) additionally required that, with some large probability, the matched record from the additional information does not belong to another element of the released data. As such, they suggest framing the problem as a discriminant analysis for linking records, and the proposed measure of privacy is the percentage of records which were threatened by identification. They review slicing, micro-aggregates, and recombinations while noting that the addition of random noise does little to protect confidentiality in the framework.

They conclude that microdata should be released with only a few variables making it difficult or impossible for an intruder to link records. However, for

data sets with a large number of variables, it becomes very easy to create a privacy breach. As such they suggest that the only way to protect data with a large number of variables is through "massive modifications of the data" (Paass, 1988) which leads to reduced utility in analysis of the perturbed data, especially for more complicated statistical techniques.

Duncan and Lambert (1989) proposed a measure of privacy based on decision theory. They approached the problem from the point of view that an intruder is searching for a specific target record which they refer to as t_0. After viewing the data, the possible values of the target are described by some predictive distribution, $p_y(s)$. Further, following their decision theoretic approach, they define a loss function, $L(t, s)$ where the action taken is choosing the target to be t, but, in fact, s is the actual target. Therefore, the expected loss can be found by integrating over all possible values of s, namely, $\int L(t, s)p_y(s)ds$. The t which minimizes expected loss is the best choice for the intruder. Along with this they define the uncertainty of the intruder as $U(y) = inf_t \int L(t, s)p_y(s)ds$ which is minimum expected loss. This quantity, which can be viewed as the intruders uncertainty, shows that the data are well protected when this gets large indicating more uncertainty. Reiter (2005a) uses the Duncan-Lambert framework to assess privacy of several disclosure control techniques under different assumptions of the knowledge of the intruder.

Bethlehem et al. (1990) proposed a measure of privacy that they refer to as the resolution of the "key", a set of variables used for identification. The measure is based on the uniqueness of elements in the population. Here, the term "key" is similar to the term "quasi-identifier" used in Dalenius (1986) and, later, in Sweeney (2002b). They defined the resolution as $R = (\sum_{i=1}^{K} \pi_i^2)^{-1}$ where $\pi_i = \frac{F_i}{N}$ with N being the size of the population and F_i is the number of elements in the population with key value i for $i = 1...k$.

Marsh et al. (1991) proposed a measure of quantifying privacy which uses uniqueness as a component along with other quantities. They are mainly interested in assessing privacy in a sample of anonymized records (SAR). They suggested that "One way to think of the real risk is as the total probability of an individual being identified from the SAR." As such, they set forth four conditions which the user must create to cause a privacy issue:

a. Key variables recorded identically in both data sets
b. Presence in the SAR
c. Population Uniqueness
d. Verification of population uniqueness

Using these they define a measure of privacy based on the conditional probability of an identification occurring given that an attempt at a privacy breach has taken place.

Skinner et al. (1994) proposed a measure of privacy based on, not only population uniqueness (PU), but also sample uniqueness (SU). Namely, they proposed assessing privacy as Pr(PU | SU). Previously, the probability of a PU was used to assess privacy. However, that method misses the fact that the intruder will only have access to the released sample of the data. Therefore, the intruder

can only possibly create a privacy breach for observations that are SU. So, a privacy breach occurs when a SU is verified to be PU. They outline three possible ways that a record could be verified as PU:

1. Population lists - If an individual had access to a population list of identifying features, individuals could easily be verified as unique.
2. Statistical Inference - A statistical argument can be made that certain combinations of characteristics are extremely rare and are, with some high probability, population unique.
3. Figures in the public eye - Certain combinations of characteristics will exist that the public will know all people with those characteristics. They offer, among others, the example of a police chief with 9 children and a Ph. D. This record is easily identifiable in the population. They also mention that easily recognized groups will be easy to identify, such as, if occupation is listed as "US Senator". That population is easily verified.

They note that statistical inference is the most likely way to verify population uniqueness, as the census has a good deal of control in preventing the other two. The authors go on to demonstrate how the Poisson-Gamma model of Bethlehem et al. (1990) would be used in this situation for estimating the probability of population uniqueness.

Skinner and Elliot (2002) discussed and reviewed two previous measures of privacy, including the Skinner et al. (1994) proposal. The paper then goes on to propose another measure based on Elliot (2000). The first measure of privacy is Pr(PU), and the second is the proportion of records which are both PU and SU to the number of records which are SU. The new measure of privacy, which Skinner and Elliot (2002) refer to as Θ is

$$\Theta = \frac{\Sigma_j I(f_j = 1)}{\Sigma_j F_j I(f_j = 1)}$$

where f_j and F_j are, respectively, the frequency of the j-th combination of identifying features in the sample and the frequency of the j-th combination of identifying features in the population. This measure can be thought of as the probability of a correct match given the probability of a unique match.

Skinner and Elliot (2002) go on to review these three measures. They argue that the first measure, $Pr(PU)$, which is the probability that an observation is PU, is overly optimistic. So they go on to compare the second measure, $Pr(PU|SU)$, which measures the probability that an observation is PU given that it is SU. They argue that their proposed measure, Θ, is an improvement over these measures.

Skinner and Shlomo (2008) investigated privacy measures which are based on the risk of re-identification. The measures they were interested in involve f_j and F_j which are, as before, respectively, the frequency of the j-th combination of identifying features in the sample and the frequency of the j-th combination of identifying features in the population. In a practical setting, since the f_j are observed but the F_j are not, the F_j must be estimated based on the observed

f_j's. F_j can be estimated using log-linear models. Thus, they break their general approach down into three steps:

1. Specifying the 'key' variables
2. Selecting one or more log-linear models which fit well according to the diagnostic criteria developed in Skinner and Shlomo (2008)
3. Use the well-fitting models to obtain risk estimates.

Here, the 'key' variables mentioned in step 1 are the same 'key' variables described in Bethlehem et al. (1990) which are variables that could be used for identification.

Sweeney (2002b) described a measure of privacy called k-anonymity, similar to methods described in Dalenius (1986). This measure is based on what Dalenius (1986) terms a quasi-identifer, which is a set of attributes in a data set that could be used for matching with an external database.

Dalenius (1986) mentioned two situations, the first when an individual has a unique set of identifiers and, the second, when only a small number of individuals has a specific set of identifiers. In the first case, it is easy to identify the individual. In the second, it would be possible to identify an individual through collusion. For example, if k individuals have a specific set of identifiers, $k-1$ individuals who have a specific trait could get together and accurately identify the remaining individual.

Sweeney (2002b) offered a real world example of a privacy breach using publicly available voting records and de-identified insurance data. Based on the unique combination of attributes from the voting records, she accurately identified former Massachusetts' governor William Weld's released insurance data. This matching was easy, as the ex-governor's combination of attributes was unique in the population. Therefore, to improve privacy, a table with multiple observations for all observed combinations of quasi-identifiers is desirable.

Thus, k-anonymity is achieved in a table if for each combination of a quasi-identifer for that table, the quasi-identifier combination appears at least k times in the table. Examples are shown in tables 2A and 2B which achieves 3-anonymity with quasi-identifier gender and race.

Sweeney (2002a) described a procedure for achieving this level of security via generalization and suppression. Generalization is achieved by grouping a possibly identifying procedure into a broader category. For example, rather than report the town or city someone lives in, a more general grouping can be achieved by reporting only the state of residence. Suppression is simply not releasing a sensitive value. By using both generalization and suppression, k-anonymity can be achieved for any data set. Two algorithms, Data fly (Sweeney, 1997) and μ-Argus (Hundepool and Willenborg, 1996, Hundepool et al., Feb. 2005), can be used to anonymize data using this approach, however, Sweeney (2002a, Page 12) notes that "Datafly can over distort data and μ-Argus can additionally fail to provide adequate protection.".

Machanavajjhala et al. (2007) described the shortcomings of k-anonymity by explaining how privacy can still be breached even when k-anonymity is achieved.

TABLE 2
(A) 3-anonymous data (B) 3-anonymous data, but disclosures take place

(A) Cancer Data			(B) Cancer Data		
Gender	Race	Type of Cancer	Gender	Race	Type of Cancer
M	White	Throat	M	White	Prostate
M	White	Prostate	M	White	Prostate
M	White	Lung	M	White	Prostate
M	Black	Lung	M	Black	Lung
M	Black	Prostate	M	Black	Prostate
M	Black	Lung	M	Black	Lung
M	Black	Stomach	M	Black	Stomach
F	White	Breast	F	White	Breast
F	White	Lung	F	White	Lung
F	White	Breast	F	White	Breast

Two types of attacks on privacy are mentioned. The first is an attack when the sensitive attributes lack diversity. Table 2B achieves 3-anonymity, but here all white males suffer from prostate cancer. Thus a disclosure has taken place because one can now infer that each specific white male in the database has prostate cancer. Note that an identity disclosure has not taken place here. Our target may be a white male and our goal as an intruder is to find out what type of cancer they have. While we are unable to match to a particular record in the database, we can still discover that our target must have prostate cancer (provided we know our target is in the database).

Another type of attack discussed in Machanavajjhala et al. (2007) is one when the intruder has background information. This could occur if, for example, he or she knew the identity of the white female in this database with lung cancer. Using the data in table 2B together with this background information allows one to conclude that each of the other white females in this database must have breast cancer. Again, we cannot match the records of the white females with breast cancer to specific individuals, but the sensitive attribute is still disclosed.

In response to the problems presented from k-anonymity, Machanavajjhala et al. (2007) proposed l-diversity. l-diversity ensures that within each equivalence class, that the values of the sensitive attributes are all "well represented". Tables 3A and 3B both demonstrate a 3-diverse table.

Li et al. (2007) discussed the weaknesses of l-diversity and proposes t-closeness as an alternative. Li et al. (2007) mentions two types of attacks that are possible even when l-diversity is achieved. The first, a skewness attack, occurs when the distribution of the sensitive attributes differs significantly from that of the overall population. For example, if some disease in a population is rare, but the prevalence of a disease within an equivalence class is much higher. In this scenario, an intruder has gained some sensitive information about a group of people. The second type of attack is a similarity attack. This occurs when sensitive attributes are technically different, but similar in nature. Li et al. (2007) offered an example where an intruder is able to infer that an individual has

TABLE 3
(A) 3-anonymous and 3-diverse data (B) 3-anonymous and 3-diverse data, but disclosures take place

(A) Cancer Data			(B) Disease Data		
Gender	Race	Type of Cancer	Gender	Race	Disease
M	White	Prostate	M	White	Prostate Cancer
M	White	Lung	M	White	Lung Cancer
M	White	Stomach	M	White	Gastritis
M	Black	Lung	M	Black	Lung Cancer
M	Black	Prostate	M	Black	Gastric Ulcer
M	Black	Lung	M	Black	Lung Cancer
M	Black	Stomach	M	Black	Gastritis
F	White	Breast	F	White	Stomach Cancer
F	White	Lung	F	White	Gastric Ulcer
F	White	Stomach	F	White	Gastritis

either gastritis, gastric ulcer, or stomach cancer. Table 3B shows an example of a table that achieves 3-anonymity and 3-diversity. However, while an intruder cannot find out exactly what disease a white female has, the intruder can still infer that the individual has a medical problem related to the stomach. Thus a sensitive attribute has been disclosed.

Therefore, Li et al. (2007) defined t-closeness as follows: (The t-closeness Principle): An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

While there are many different choices for assessing the distance described in the definition of t-closeness, Li et al. (2007) settled on the Earth Mover's distance (Rubner et al., 1998), a metric for comparing the difference between two distributions, for its desirable properties in this situation.

3.2. Inferential and predictive disclosure measures

Willenborg and de Waal (2001, 42-43) says "...disclosure was said to occur if the intruder is able to use the microdata to gain information about the target unit. In loose terms, this defines predictive disclosure." Willenborg and de Waal (2001, Page 43) then offers a simple example of predictive disclosure. Suppose a released microdata set contains information about individuals gender, age, occupation, and region as well as income. Now suppose that an intruder is interested in a target about which they know the gender, age, occupation and region, but they do not know income. The intruder can build a regression model based on the microdata with income as the response variable and gender, age, occupation, and region as predictors. Using this regression model, an intruder now has a predictive distribution of the target's income.

Dwork (2006, Page 2) offered an example of inferential disclosure. Suppose that someone's height is considered to be sensitive information. Now consider

that there is a database which gives someone information about the average heights of women of different nationalities. If an intruder has access to this database and the information that "Terry Gross is two inches shorter that the average Lithuanian woman", they now know the exact height of Terry Gross.

Note that in the last two examples, the target does not need to be part of the released information for a privacy breach to occur. This is quite a startling result. Not only does one need to protect the privacy of the individuals in the released microdata, one may also need to make privacy considerations for individuals who are not in the released database!

3.2.1. Knowledge, Knowledge Gain, and Relative Knowledge Gain

Duncan and Lambert (1986, Page 13) proposed three types of measures of disclosure risk all based on the following two principles:

1. The complete state of a user's uncertainty about a target before and after data release is specified by the user's prior and posterior predictive distributions, respectively.
2. The user's uncertainty about a target can be summarized by applying a nonnegative concave function

Let $U(\cdot)$ be an uncertainty function with larger values of U indicating more uncertainty (DeGroot, 1962, 1970). Therefore, privacy can be measured based on the difference between the uncertainty in the prior predictive distribution ($U(prior)$) and the uncertainty in the updated posterior predictive distribution ($U(posterior)$). Thus, Duncan and Lambert (1986) defined knowledge to be $U(posterior)$, knowledge gain to be $U(posterior) - U(prior)$, and relative knowledge gain to be $\frac{U(posterior)-U(prior)}{U(prior)}$. Using one of these measures, data would not be released if the measure exceeded a pre-specified threshold.

One drawback to this method is that a data releasing agency must specify the prior beliefs of a potential consumer of the data. It is often impossible to know exactly what potential data a user has, complicating the specification of a user's prior. An agency will be able to more accurately define the prior beliefs of data user if they have some idea of what data sets are already available to a data consumer. Therefore, it may be useful for releasing agencies to keep records of what other data are readily available to a potential consumer of the data.

Finally, Duncan and Lambert (1986, Page 17) noted that "Although specification of uncertainty functions and disclosure limits may appear arbitrary and difficult to justify, it is also difficult to justify rigorously ad hoc rules for releasing data." This statement emphasizes the need for an easily interpretable, reasoned metric for assessing privacy, with which a set of guidelines could be produced to direct data releasing agencies in the proper manner in which to release data to the public while maintaining sufficient levels of privacy.

3.2.2. Differential privacy

Ideally, as proposed in Dalenius (1977), "access to a statistical database should not enable one to learn anything about an individual that could not be learned without access to the database." However, Dwork (2006) showed that this level of privacy is not achievable. Therefore, Dwork (2006) proposed differential privacy which measures the added risk of disclosure when an individual decides to participate in a database. This type of privacy moves away from absolute guarantees. Instead, differential privacy offers relative guarantees of privacy. This type of privacy guarantee ensures that the risk of disclosure for a given individual is nearly the same whether or not an individual is in a given database.

Consider a database D with n observations, and a neighboring database D' which differs from D by at most one observation.

Dwork (2006) then defined differential privacy as follows: A randomized function κ_f has ϵ-differential privacy if $Pr[\kappa_f(D) \in S] \leq e^{\epsilon} Pr[\kappa_f(D') \in S]$ for any two databases D and D' that differ by at most one element and all $S \subseteqq range(\kappa)$. This definition can be extended to group privacy for g persons by replacing e^{ϵ} with $e^{g\epsilon}$.

As an example, consider that the query of interest is the mean ($\bar{X}$) of the database D. Rather than releasing the true response to this query, a randomized version of the query is released. This could be the result of the query with the addition of Laplace noise. Therefore, the released values is simply a draw from a Laplace distribution with mean $\bar{X}$ and scale parameter σ. Dwork (2006) showed that this type of randomized release mechanism achieves $\frac{\Delta f}{\sigma}$-differential privacy where $\Delta f = Max_{D,D'}||f(D) - f(D')||_1$ and is called the L_1 sensitivity.

A relaxed form ϵ-differential privacy, proposed in Nissim et al. (2007), is (ϵ, δ)-indistinguishability. This is necessary, as some random release mechanisms cannot achieve ϵ-differential privacy. Again let κ be a randomized function of the data D and S be the domain of $\kappa(D)$. Then for $\epsilon > 0$ and data sets (D, D') differing by exactly one row, $\kappa(D)$ achieves (ϵ, δ)-indistinguishability if for any subset $S \subseteqq range(\kappa)$ $Pr(\kappa(D) \in S) \leq e^{\epsilon} Pr(\kappa(D') \in S) + \delta$.

Machanavajjhala et al. (2008) proposed a variant of differential privacy called probabilistic differential privacy. This method is similar to ϵ-differential privacy, however, the authors argued that certain events are so rare that they should not be used in the calculation of privacy. Thus, their proposed metric considers only events that are not extremely rare.

The authors offered an example of this using data on commuters. The origins and destinations are captured as k blocks, and there is a histogram for the destination block. The histogram for any block d can be thought of as a vector $(n_1, n_2, ..., n_k)$ where n_i is the number of persons that commute from block i to block d (for $1 \leq i \leq k$).

Synthetic data in this case corresponding to block d are simply a sample of size m from a multinomial distribution with a Dirichlet prior whose parameters are $(n_1 + \alpha_1, n_2 + \alpha_2, ..., n_k + \alpha_k)$. Letting $\vec{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k)$ one can calculate the differential privacy by computing the maximum value

of $\frac{Pr((m_1,m_2,...,m_k)|(n_1,n_2,...,n_k),\vec{\alpha})}{Pr((m_1,m_2,...,m_k)|(n'_1,n'_2,...,n'_k),\vec{\alpha})}$. Here, $(n'_1, n'_2, ..., n'_k)$ is the histogram corresponding to D' such that D and D' differ by one observation.

If D and D' differ in block i, the authors show that, the worst case value of the above ratio happens when all m points in the synthetic data fall into block d. However, the probability of this happening is extremely small. In this case, the synthetic data are highly unrepresentative of the original data. The idea of probabilistic differential privacy is to compute the value of ϵ using a probable value of the above ratio, rather than the worst case maximum value. As one can see, the probabilistic differential privacy is closely related to differential privacy.

Additional related work includes Dwork and Nissam (2004), Dinur and Nissam (2003), Blum et al. (2005), Dwork et al. (2006), Smith (2008), Dwork (2008), Dwork and Lei (2009), and Wasserman and Zhou (2010).

3.2.3. Using the Reciever-Operator Character (ROC) Curve to assess privacy

Matthews et al. (2010a) discussed the assessment of privacy in a hypothesis testing framework. As in differential privacy, consider a database D and a neighboring database D', as well as, a query of interest, $f : D \to R^d$. However, rather than release the results of the query, some noise is added via a randomized function κ, yielding $\kappa_f(D)$, which is the returned randomized result of query f. Differential privacy attempts to bound this ratio $\frac{Pr[\kappa_f(D)\in S]}{Pr[\kappa_f(D')\in S]}$ for all D, D' and $S \subseteq range(\kappa)$. If this ratio can be bounded by a small value, that indicates that the two distribution do not differ greatly, whereas if the ratio is bounded only by a large value, this indicates that the two distributions are much different.

In another light, this problem can be viewed as a hypothesis test of $H_0 : f(D) = f(D')$ vs $H_1 : f(D) \neq f(D')$ where we view the database D and D' as a population about which an intruder wishes to make inference. This is done since an intruder is only interested in making inferences about the data (whereas a researcher strives to make inferences about the population). Further, the released randomized version of the query, $\kappa_f(D)$ and $\kappa_f(D')$, are released as data. Then one can use the ratio $\frac{Pr[\kappa_f(D)\in S]}{Pr[\kappa_f(D')\in S]}$ as the basis for a likelihood ratio test statistic to test the hypothesis of interest. Therefore, in this framework, the ϵ in ϵ-differential privacy can be viewed as a test statistics for comparing the distributions generated by D and D'. However, in a hypothesis test, one should not only consider the value of the test statistic, but also the distribution of the test statistic.

Now, one can study the relationship between the sensitivity and specificity of this test by plotting specificity versus sensitivity to create an ROC curve. Next, one can consider the area under the curve (AUC), which can range from .5 to 1. When this area is near 1, that indicates a very good test, which in this case corresponds to low levels of privacy. This results when the randomized release function based on D' is significantly different than the one based on D'. However, when the AUC is near .5, this indicates a poor test implying high

privacy. This occurs when the randomized release function based on D does not change very much when we based the same function on D'.

Thus, we can define risk as the maximum value of the AUC over all possible neighboring databases D and D': $Risk = \max_{D,D'} AUC$. Similarly, we can define $Privacy = 1 - Risk$ which ranges from 0, low privacy, to .5, high privacy.

3.2.4. Model Disclosure

Palley and Simonoff (1987) demonstrated how privacy in a database can be breached by using regression to infer the values of confidential attributes relying on data from simple queries to the database. They used R^2 as a measure of their predictive accuracy. Thus $1 - R^2$ can be viewed as the level of privacy. However, simply because one confidential attribute is private does not mean that all confidential attributes are private. This is discussed in Tendick (1991).

Sarathy and Muralidhar (2002b) proposed the use of canonical correlation analysis as a technique for assessing privacy. Canonical correlation analysis was used, in general, to find and quantify relationships between two groups of variables. In a privacy setting, one is interested in the relationship between the variables with no privacy restrictions and the variables that are confidential so canonical correlation analysis lends itself in some natural way to the assessment of privacy.

3.2.5. Other Methods

Rajasekaran et al. (2009) proposed a measure of privacy based on the idea that observations far from the mean will be easily identified. However, rather than measuring an observation's distance from the overall mean, the data are first clustered, and observations which are far away from the center of the respective cluster are considered to be at high risk for disclosure. Further, Rajasekaran et al. (2009) also suggests that if observations fall into clusters with too few observations, they should be suppressed as they are observations at high risk of disclosure.

4. Summary

Statistical disclosure limitation is a very broad topic. However, many areas of research depend on data that can only be used if privacy is maintained thus highlighting the importance of disclosure limitation. Perfect privacy, never releasing any confidential data, and perfect utility, releasing all confidential data, both present significant problems. Therefore, some balance must be struck between these competing goals.

This paper offers a summary of methods which have been proposed to maintain the privacy of the individual in public release data sets, as well as, a review of proposed techniques for assessing the privacy of some of the privacy preserving methods.

Clearly, there will be some trade off between the amount of privacy ensured and the utility of the released data. Many privacy preserving techniques work by perturbing the data to be released, resulting in potentially less useful data. Karr et al. (2006) formally discussed the risk utility trade-off for several methods of statistical disclosure techniques including addition of noise, rank swapping, microaggregation, and resampling. Other references which discuss this trade-off include Domingo-Ferrer and Torra (2001b,a) and Kaufman et al. (2005).

While many methods of preserving privacy have been proposed, there are not, as of yet, any formal guidelines for many data releasing institutions to follow when releasing data to the public (although attempts have been made (United Nations Economic Comission for Europe (UNECE), 2007, Hundepool et al., 2006, Federal Committee on Statistical Methodology (FCSM), 2005)). Most data releases are deemed to be "private enough" with no formal assessment of the privacy ensured. Eventually, legal guidelines will need to be set as to what constitutes adequate amounts of privacy. This accentuates the need for an easily interpretable measure of privacy that can be used by policy makers in drafting legislation pertaining to adequate privacy levels for public release data sets.

While this paper is by no means an exhaustive reference on statistical disclosure limitation, the hope is that this manuscript will provide an introduction to disclosure limitation techniques, as well as, methods for measuring privacy.

The goal of statistical disclosure limitation efforts is to make sure that data are used for research, rather than malicious purposes, including the disclosure of individuals' private information. More work is needed in both areas, the development of statistical disclosure control techniques and the assessment of privacy, as the importance of this field becomes increasingly relevant as we continue on into an age ruled by data.

References

Abowd, J., Woodcock, S., 2001. Disclosure limitation in longitudinal linked data. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, 215–277.

Adam, N.R., Worthmann, J.C., 1989. Security-control methods for statistical databases: a comparative study. ACM Comput. Surv. 21 (4), 515–556.

Armstrong, M., Rushton, G., Zimmerman, D.L., 1999. Geographically masking health data to preserve confidentiality. Statistics in Medicine 18 (5), 497–525.

Bethlehem, J.G., Keller, W., Pannekoek, J., 1990. Disclosure control of microdata. Jorunal of the American Statistical Association 85, 38–45.

Blum, A., Dwork, C., McSherry, F., Nissam, K., 2005. Practical privacy: The sulq framework. In: Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. pp. 128–138.

Bowden, R.J., Sim, A.B., 1992. The privacy bootstrap. Journal of Business and Economic Statistics 10 (3), 337–345.

Carlson, M., Salabasis, M., 2002. A data-swapping technique for generating synthetic samples; a method for disclosure control. Res. Official Statist. (5), 35–64.

Cox, L.H., 1980. Suppression methodology and statistical disclosure control. Journal of the American Statistical Association 75, 377–385.

Cox, L.H., 1984. Disclosure control methods for frequency count data. Tech. rep., U.S. Bureau of the Census.

Cox, L.H., 1987. A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association 82, 520–524.

Cox, L.H., 1994. Matrix masking methods for disclosure limitation in microdata. Survey Methodology 6, 165–169.

Cox, L.H., Fagan, J.T., Greenberg, B., Hemmig, R., 1987. Disclosure avoidance techniques for tabular data. Tech. rep., U.S. Bureau of the Census.

Dalenius, T., 1977. Towards a methodology for statistical disclosure control. Statistik Tidskrift 15, 429–444.

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

Dalenius, T., Denning, D., 1982. A hybrid scheme for release of statistics. Statistisk Tidskrift.

Dalenius, T., Reiss, S.P., 1982. Data-swapping: A technique for disclosure control. Journal of Statistical Planning and Inference 6, 73–85. MR0653248

De Waal, A., Hundepool, A., Willenborg, L., 1995. Argus: Software for statistical disclosure control of microdata. U.S. Census Bureau.

DeGroot, M.H., 1962. Uncertainty, information, and sequential experiments. Annals of Mathematical Statistics 33, 404–419. MR0139242

DeGroot, M.H., 1970. Optimal Statistical Decisions. Mansell, London. MR0356303

Dinur, I., Nissam, K., 2003. Revealing information while preserving privacy. In: Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principlesof Database Systems. pp. 202–210.

Domingo-Ferrer, J., Torra, V., 2001a. A Quantitative Comparison of Disclosure Control Methods for Microdata. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (Eds.), Confidentiality, Disclosure and Data Access - Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam, Ch. 6, pp. 113–135.

Domingo-Ferrer, J., Torra, V., 2001b. Disclosure control methods and information loss for microdata. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (Eds.), Confidentiality, Disclosure and Data Access - Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam, Ch. 5, pp. 93–112.

Duncan, G., Lambert, D., 1986. Disclosure-limited data dissemination. Journal of the American Statistical Association 81, 10–28.

Duncan, G., Lambert, D., 1989. The risk of disclosure for microdata. Journal of Business & Economic Statistics 7, 207–217.

Duncan, G., Pearson, R., 1991. Enhancing access to microdata while protecting confidentiality: prospects for the future (with discussion). Statistical Science 6, 219–232.

Dwork, C., 2006. Differential privacy. In: ICALP. Springer, pp. 1–12. MR2307219

Dwork, C., 2008. An ad omnia approach to defining and achieving private data analysis. In: Lecture Notes in Computer Science. Springer, p. 10. MR2581844

Dwork, C., Lei, J., 2009. Differential privacy and robust statistics. In: Proceedings of the 41th Annual ACM Symposium on Theory of Computing (STOC). pp. 371–380.

Dwork, C., Mcsherry, F., Nissim, K., Smith, A., 2006. Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Theory of Cryptography Conference. Springer, pp. 265–284. MR2241676

Dwork, C., Nissam, K., 2004. Privacy-preserving datamining on vertically partitioned databases. In: Advances in Cryptology: Proceedings of Crypto. pp. 528–544. MR2147523

Elliot, M., 2000. DIS: a new approach to the measurement of statistical disclosure risk. International Journal of Risk Assessment and Management 2, 39–48.

Federal Committee on Statistical Methodology (FCSM), 2005. Statistical policy working group 22 - report on statistical disclosure limitation methodology. U.S. Census Bureau.

Fellegi, I.P., 1972. On the question of statistical confidentiality. Journal of the American Statistical Association 67 (337), 7–18.

Fienberg, S.E., McIntyre, J., 2004. Data swapping: Variations on a theme by Dalenius and Reiss. In: Domingo-Ferrer, J., Torra, V. (Eds.), Privacy in Statistical Databases. Vol. 3050 of Lecture Notes in Computer Science. Springer Berlin/Heidelberg, pp. 519, http://dx.doi.org/10.1007/978-3-540-25955-8_2

Fuller, W., 1993. Masking procedurse for microdata disclosure limitation. Journal of Official Statistics 9, 383–406.

General Assembly of the United Nations, 1948. Universal declaration of human rights.

Gouweleeuw, J., P. Kooiman, L.W., de Wolf, P.-P., 1998. Post randomisation for statistical disclosure control: Theory and implementation. Journal of Official Statistics 14 (4), 463–478.

Greenberg, B., 1987. Rank swapping for masking ordinal microdata. Tech. rep., U.S. Bureau of the Census (unpublished manuscript), Suitland, Maryland, USA.

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R., Horvitz, D.G., 1969. The unrelated question randomized response model: Theoretical framework. Journal of the American Statistical Association 64 (326), 520–539. MR0247719

Harel, O., Zhou, X.-H., 2007. Multiple imputation: Review and theory, implementation and software. Statistics in Medicine 26, 3057–3077. MR2380504

Hundepool, A., Domingo-ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E.S., Seri, G., paul De Wolf, P., 2006. A CENtre of EXcellence for Statistical Disclosure Control Handbook on Statistical Disclosure Control Version 1.01.

Hundepool, A., Wetering, A. v.d., Ramaswamy, R., Wolf, P.d., Giessing, S., Fischetti, M., Salazar, J., Castro, J., Lowthian, P., Feb. 2005. τ-argus 3.1 user manual. Statistics Netherlands, Voorburg NL.

Hundepool, A., Willenborg, L., 1996. μ- and τ-argus: Software for statistical disclosure control. Third International Seminar on Statistical Confidentiality, Bled.

Karr, A., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P., 2006. A framework for evaluating the utility of data altered to protect confidentiality. American Statistician 60 (3), 224–232. MR2246755

Kaufman, S., Seastrom, M., Roey, S., 2005. Do disclosure controls to protect confidentiality degrade the quality of the data? In: American Statistical Association, Proceedings of the Section on Survey Research.

Kennickell, A.B., 1997. Multiple imputation and disclosure protection: the case of the 1995 survey of consumer finances. Record Linkage Techniques, 248–267.

Kim, J., 1986. Limiting disclosure in microdata based on random noise and transformation. Bureau of the Census.

Krumm, J., 2007. Inference attacks on location tracks. Proceedings of Fifth International Conference on Pervasive Computingy, 127–143.

Li, N., Li, T., Venkatasubramanian, S., 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. pp. 106–115.

Liew, C.K., Choi, U.J., Liew, C.J., 1985. A data distortion by probability distribution. ACM Trans. Database Syst. 10 (3), 395–411.

Little, R.J.A., 1993. Statistical analysis of masked data. Journal of Official Statistics 9, 407–426.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. John Wiley & Sons. MR0890519

Liu, F., Little, R.J.A., 2002. Selective multiple mputation of keys for statistical disclosure control in microdata. In: Proceedings Joint Statistical Meet. pp. 2133–2138.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L., April 2008. Privacy: Theory meets practice on the map. In: International Conference on Data Engineering. Cornell University Comuputer Science Department, Cornell, USA, p. 10.

Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M., 2007. L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1 (1), 3.

Manning, A.M., Haglin, D.J., Keane, J.A., 2008. A recursive search algorithm for statistical disclosure assessment. Data Min. Knowl. Discov. 16 (2), 165–196. MR2412605

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N., 1991. The case for samples of anonymized records from the 1991 census. Journal of the Royal Statistical Society 154 (2), 305–340.

Matthews, G.J., Harel, O., Aseltine, R.H., 2010a. Assessing database privacy using the area under the receiver-operator characteristic curve. Health Services and Outcomes Research Methodology 10 (1), 1–15.

Matthews, G.J., Harel, O., Aseltine, R.H., 2010b. Examining the robustness of fully synthetic data techniques for data with binary variables. Journal of Statistical Computation and Simulation 80 (6), 609–624.

Moore, Jr., R., 1996. Controlled data-swapping techniques for masking public use microdata. Census Tech Report.

Mugge, R., 1983. Issues in protecting confidentiality in national health statistics. Proceedings of the Section on Survey Research Methods.

Nissim, K., Raskhodnikova, S., Smith, A., 2007. Smooth sensitivity and sampling in private data analysis. In: STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing. pp. 75–84. MR2402430

Paass, G., 1988. Disclosure risk and disclosure avoidance for microdata. Journal of Business and Economic Statistics 6 (4), 487–500.

Palley, M., Simonoff, J., 1987. The use of regression methodology for the compromise of confidential information in statistical databases. ACM Trans. Database Systems 12 (4), 593–608.

Raghunathan, T.E., Reiter, J.P., Rubin, D.B., 2003. Multiple imputation for statistical disclosure limitation. Journal of Official Statistics 19 (1), 1–16.

Rajasekaran, S., Harel, O., Zuba, M., Matthews, G.J., Aseltine, Jr., R., 2009. Responsible data releases. In: Proceedings 9th Industrial Conference on Data Mining (ICDM). Springer LNCS, pp. 388–400.

Reiss, S.P., 1984. Practical data-swapping: The first steps. CM Transactions on Database Systems 9, 20–37.

Reiter, J.P., 2002. Satisfying disclosure restriction with synthetic data sets. Journal of Official Statistics 18 (4), 531–543.

Reiter, J.P., 2003. Inference for partially synthetic, public use microdata sets. Survey Methodology 29 (2), 181–188.

Reiter, J.P., 2004a. New approaches to data dissemination: A glimpse into the future (?). Chance 17 (3), 11–15. MR2061931

Reiter, J.P., 2004b. Simultaneous use of multiple imputation for missing data and disclosure limitation. Survey Methodology 30 (2), 235–242.

Reiter, J.P., 2005a. Estimating risks of identification disclosure in microdata. Journal of the American Statistical Association 100, 1103–1112. MR2236926

Reiter, J.P., 2005b. Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society, Series A: Statistics in Society 168 (1), 185–205. MR2113234

Reiter, J.P., 2005c. Using CART to generate partially synthetic public use microdata. Journal of Official Statistics 21 (3), 441–462.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons. MR0899519

Rubin, D.B., 1993. Comment on "Statistical disclosure limitation". Journal of Official Statistics 9, 461–468.

Rubner, Y., Tomasi, C., Guibas, L.J., 1998. A metric for distributions with applications to image databases. Computer Vision, IEEE International Conference on 0, 59.

Sarathy, R., Muralidhar, K., 2002a. The security of confidential numerical data in databases. Information Systems Research 13 (4), 389–403.

Sarathy, R., Muralidhar, K., 2002b. The security of confidential numerical data in databases. Info. Sys. Research 13 (4), 389–403.

Schafer, J.L., Graham, J.W., 2002. Missing data: Our view of state of the art. Psychological Methods 7 (2), 147–177.

Singh, A., Yu, F., Dunteman, G., 2003. MASSC: A new data mask for limiting statistical information loss and disclosure. In: Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality. pp. 373–394.

Skinner, C., 2009. Statistical disclosure control for survey data. In: Pfeffermann, D and Rao, C.R. eds. Handbook of Statistics Vol. 29A: Sample Surveys: Design, Methods and Applications. pp. 381–396. MR2654645

Skinner, C., Marsh, C., Openshaw, S., Wymer, C., 1994. Disclosure control for census microdata. Journal of Official Statistics 10, 31–51.

Skinner, C., Shlomo, N., 2008. Assessing identification risk in survey microdata using log-linear models. Journal of the American Statistical Association 103, 989–1001. MR2462887

Skinner, C.J., Elliot, M.J., 2002. A measure of disclosure risk for microdata. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64 (4), 855–867. MR1979391

Smith, A., 2008. Efficient, dfferentially private point estimators. arXiv:0809.4794v1 [cs.CR].

Spruill, N.L., 1982. Measures of confidentiality. Statistics of Income and Related Administrative Record Research, 131–136.

Spruill, N.L., 1983. The confidentiality and analytic usefulness of masked business microdata. In: Proceedings of the Section on Survey Reserach Microdata. American Statistical Association, pp. 602–607.

Sweeney, L., 1996. Replacing personally-identifying information in medical records, the scrub system. In: American Medical Informatics Association. Hanley and Belfus, Inc., pp. 333–337.

Sweeney, L., 1997. Guaranteeing anonymity when sharing medical data, the datafly system. Journal of the American Medical Informatics Association 4, 51–55.

Sweeney, L., 2002a. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 10 (5), 571–588. MR1948200

Sweeney, L., 2002b. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 10 (5), 557–570. MR1948199

Tendick, P., 1991. Optimal noise addition for preserving confidentiality in multivariate data. Journal of Statistical Planning and Inference 27 (2), 341–353. MR1108554

United Nations Economic Comission for Europe (UNECE), 2007. Manging statistical cinfidentiality and microdata access: Principles and guidlinesof good practice.

Warner, S.L., 1965. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60 (309), 63–69.

Wasserman, L., Zhou, S., 2010. A statistical framework for differential privacy. Journal of the American Statistical Association 105 (489), 375–389. MR2656057

Willenborg, L., de Waal, T., 2001. Elements of Statistical Disclosure Control. Springer-Verlag. MR1866909

Woodward, B., 1995. The computer-based patient record and confidentiality. The New England Journal of Medicine, 1419–1422.