Measuring Disclosure Risk for a Synthetic Data Set Created Using Multiple Methods

Jennifer C. Huckett Michael D. Larsen*

Abstract

Government agencies must simultaneously maintain confidentiality of individual records and disseminate useful microdata. We propose a method to create synthetic data that combines quantile regression, hot deck imputation, and rank swapping. The result from implementation of the proposed procedure is a releasable data set containing original values for a few key variables, synthetic quantile regression predictions for several variables, and imputed and perturbed values for remaining variables. To measure the disclosure risk in the resulting synthetic data set, we extend existing probabilistic risk measures that aim to imitate an intruder attempting to match a record in the released data with information previously available on a target respondent.

Key Words: hot deck imputation, quantile regression, rank swapping, statistical disclosure limitation, synthetic data

1. Introduction

Government agencies face demands to release accurate, timely data and to simultaneously uphold their promises of privacy and confidentiality to respondents. We study options for generating synthetic data files for public release. Specifically, we study combining quantile regression and hot deck imputation with rank swapping to produce releasable, usable synthetic microdata. To capture the complex relationships found in demographic and economic data collected by statistical agencies, conditional quantile regression models are used. Predicted values computed from model estimates and key predictors are generated for several confidential variables at random quantiles. Values for other variables are imputed from the original data using hot deck imputation and further perturbed using a rank swapping procedure. The quantile regression predictions are combined with the imputed perturbed values to form a data set with record level data for release that has low disclosure risk and high data utility. Details of the new procedure are described in Huckett and Larsen (2007, 2008).

In general, there is a trade off between reducing disclosure risk and increasing data utility. At one extreme, releasing no data has zero risk (except for someone physically stealing the data), but no usefulness at all. At the other extreme, releasing all collected data, including personal identifying information (or at least everything but explicit personal identifiers), should be the most useful for researchers, but has the highest potential for harm to respondents. In order to judge the relative merits of disclosure limitation methods, we need to assess both the risk of disclosing confidential information and the data utility, or the inferential worthiness, of the released data set. Released data sets that carry too much disclosure risk or too little data utility should be avoided. Details about the framework developed in Duncan and Lambert (1986, 1989) and Reiter (2005) are presented. We extend this framework to develop a disclosure risk measure for a synthetic data set generated using our proposed procedure.

2. Measuring Disclosure Risk: Introduction and Notation

Suppose the original data set is called Y, with variables $Y_0, \cdots, Y_d$ and the released (perturbed) data set is called Z, with variables $Z_1, \cdots, Z_d$. We assume an intruder with access to the released data will attempt to link one or several target records from Y with records in Z using information available from external sources on the target, t. The intruder is assumed to compute the probability that record j in the released data set belongs to the target, conditional on the information in t and Z, denoted $Pr(J = j|t, Z)$. The larger value of this probability for record j, the more likely the intruder will identify record j with the target.

The original data set contains records $j = 1, 2, ..., n$, with data on variables, $k = 0, 1, ..., d$. The agency may release $r \leq n$ records with values on all d variables or a subset thereof. Directly identifying information such as name, exact address, or social security number is recorded on variables $k = 0$. No version of these variables is released. Remaining variables, $k = 1, ..., d$, are divided into available and unavailable variables, denoted A and U, respectively. Variables in A contain information available to the intruder from outside sources while variables in U contain information unavailable to an intruder except from the released data. Assumptions about which variables are in each set can be varied, allowing an agency the flexibility to consider an intruder with detailed and accurate information on all of the variables as well as an intruder with relatively little information on any variable. Both unavailable and available variables can be further divided into variables that are perturbed p and variables that do not get perturbed d before being released by the agency. This division allows us to incorporate information the intruder is assumed to have about the SDL method used. Further, variables in the released data set that the intruder cannot know, or match to the original data, with 100% certainty belong to the set C, where $C = (Ap, U)$, or the available variables that were perturbed and all of the unavailable variables. All variables in C have been perturbed and/or are unknown to the intruder before data are released from the agency.

Data on in the original data set Y for variable k on record j is denoted y_{kj}. The notation y_j^A is used to denote original data on available variables and y_j^{Ap} to denote data on available perturbed variables in the j^{th} record. Similar notation is

[*] Iowa State University, Snedecor Hall, Ames, IA 50011, jhuckett@gmail.com, larsen@iastate.edu

used for variables in Ad, U, and C, as well as for data in the released data set Z and the target's data t. Properties of some variables are implied by the definitions. For example, $t^A = y^A$ for all records, since information on the target is assumed to include original data on available variables. Also, $t^{Ad} = y^{Ad} = z^{Ad}$ for all records since variables in the set Ad do not get perturbed. However, even though $t^{Ap} = y^{Ap}$, t^{Ap} does not necessarily equal z^{Ap} since variables in the set Ap are perturbed from their original values. This notation is used to clearly describe the probabilistic framework presented and implemented by Duncan and Lambert (1986, 1989). The components of the probability of identification are arrived at using Bayes' rule and properties of marginal, joint, and conditional distributions in Reiter (2005).

3. Component Formulation

In order to compute $Pr(J = j|t, Z)$, Reiter (2005) breaks the probability into manageable components. Each component corresponds to properties of the variables–whether they are perturbed (p) or do not get perturbed (d), whether they are available (A) or unavailable (U), and whether they are variables with known values (Ad) or values the intruder cannot know with certainty ($C = (Ap, U)$) after release. Various assumptions about the properties of each variable and what the intruder knows prior to data being released cana be incorporated in the corresponding component. Their details are discussed here.

Using Bayes' rule, Reiter (2005) expresses the probability of identification $Pr(J = j|t, Z)$ as

$$Pr(J = j|t, Z^{Ad}, Z^C) = \frac{Pr(J = j, Z^C|t, Z^{Ad})}{Pr(Z^C|t, Z^{Ad})} = \frac{Pr(Z^C|J = j, t, Z^{Ad})Pr(J = j|t, Z^{Ad})}{\sum_{j=1}^{r+1} Pr(Z^C|J = j, t, Z^{Ad})Pr(J = j|t, Z^{Ad})}. \tag{1}$$

This expression of the probability allows us to assess disclosure risk by considering various levels of intruder knowledge and behavior as well as what SDL method was used. Assessment of each component that makes up the terms in the numerator and denominator of 1 is discussed in the following sections.

3.1 Component $Pr(J = j|t, Z^{Ad})$

Recalll that the variables in Z^{Ad} are do not get perturbed prior to data being released. This implies $Z^{Ad} = Y^{Ad}$ for all records. Thus, any record in Z with values on available variables that do not get perturbed with $z^{Ad} = t^{Ad}$ could be identified as the target's given only the information the intruder has on the target and the values on the variables in Ad. If $n_t = \#$ records in Z have $z_j^{Ad} = t^{Ad}$, then, assuming the intruder knows the target is released in Z, the chance of correctly identifying record j as the target is $1/n_t$. In other words, the probability of identifying record j as the target given the target's information t and values in Z^{Ad} is $Pr(J = j|t, Z^{Ad}) = 1/n_t$ when $z_j^{Ad} = t^{Ad}$ and zero otherwise. This result depends on the assumption that the intruder knows the target is released in Z, i.e. $Pr(J = r + 1|t, Z^{Ad}) = 0$. We assume throughout that the intruder knows the target record is released in Z, that is, $j \leq r$. This is conservative as well as computationally convenient and we assume this throughout the remainder of our discussion on disclosure risk measurement. Details for computing $Pr(J = r + 1|t, Z^{Ad}) \geq 0$ are discussed in Reiter (2005).

3.2 Component $Pr(Z^C|J = j, t, Z^{Ad})$

The probability $Pr(Z^C|J = j, t, Z^{Ad})$ is the probability of observing values on variables Z^C in the released data set given that the j^{th} record belongs to the target, the intruder's information on the target, and the values on the variables that are released unperturbed. Variables Z^C are variables that the intruder cannot know with certainty because they are available and perturbed or are unavailable. It is in this component we incorporate different assumptions of intruder knowledge with respect to the SDL method used as well as any assumptions about the joint distributions. We also incorporate assumptions about intruder behavior based on knowledge possessed.

Under the assumption that records are independent, properties of joint and marginal probabilities can be used to decompose this component further as follows (Reiter 2005): $Pr(Z^C|J = j, t, Z^{Ad}) =$

$$Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C|z_j^C, J = j, t, Z^{Ad}) \times Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad}) \times Pr(z_j^{Ap}|J = j, t, Z^{Ad}). \tag{2}$$

Each term in the right hand side of Equation 2 can be formulated according to different assumptions of intruder knowledge and behavior. These terms or components are described in the following sections.

3.2.1 $Pr(z_j^{Ap}|J = j, t, Z^{Ad})$

The component $Pr(z_j^{Ap}|J = j, t, Z^{Ad})$ is the probability of observing values on available perturbed variables in record j given that the j^{th} record belongs to the target, the information the intruder has about the target, and the information in the released data set on ovariables that are available and do not get perturbed. An intruder's knowledge about the SDL method used, the conditional distrubution of y^{Ap} and z^{Ap}, and values for t^{Ap}can be incorporated in order to formulate an expression for this component.

If the values on variables k in Ap are assumed to be independent, this probability can be formulated as the product of marginal conditional distributions over $k \in Ap$. This is the approach taken in Reiter (2005). It is appropriate for the intruder who assumes SDL methods are implemented on variables in Ap independent of one another. This approach is taken in

Reiter (2005) for a data set perturbed using traditional methods. In the proposed synthetic data method however, synthetic values are generated dependent on other variable valaues. Therefore, we approach the formulation of this component using information about the conditional manner in which values are generated.

Consider a data set Y with variables Y^{Ad} and $Y^{Ap} = (Y_1, Y_2, Y_3)$. Original values on variables Y^{Ad} are copied to the data set for release, so that $Z^{Ad} = Y^{Ad}$. Suppose synthetic values for $Z^{Ap} = (Z_1^{Ap}, Z_2^{Ap}, Z_3^{Ap})$ are the predicted values from models describing conditional relationships $Y_1|X$, $Y_2|Y_1, X$, and $Y_3|Y_2, Y_1, X$, respectively. Then the joint probability $Pr(z_j^{Ap}|J = j, t, Z^{Ad})$ can be written as the product of the sequence of marginal conditional probabilities: $Pr(z_{1,j}^{Ap}|J = j, t, Z^{Ad}) \times Pr(z_{2,j}^{Ap}|z_{1,j}^{Ap}, J = j, t, Z^{Ad}) \times Pr(z_{3,j}^{Ap}|z_{2,j}^{Ap}, z_{1,j}^{Ap}, J = j, t, Z^{Ad})$. When the models used to describe conditional relationships have known inference, the probabilities above can be estimated accordingly. This component of disclosure risk can thus be computed for intruders assumed to know the SDL procedure used. Section 3 presents details to estimate this component when quantile regression models, hot deck imputation, and rank swapping are used to generate values for Z^{Ap}.

3.2.2 $Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad})$

The component $Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad})$ is the probability of observing values on unavailable (both perturbed and not perturbed) variables in record j given the j^{th} record belongs to the target, information the intruder might have for a particular target record, the information in the released data set on variables that are available and do not get perturbed, and the value in released record j on available perturbed variables. We can incorporate an intruder's knowledge about the SDL method used and some assumed conditional distribution of y^{Ap} and z^{Ap} into formulating an expression for this component.

Consider the data set described above. Suppose synthetic values for additional variables $Y_U = (Y_4, Y_5)$ are generated using models $Y_4|Y_3, Y_2, Y_1, X$ and $Y_5|Y_4, Y_3, Y_2, Y_1, X$, respectively, producing Z_4 and Z_5. Assuming the models accurately describe the conditional distributions of Y_4 and Y_5, the joint probability $Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad})$ can be written as $Pr(z_{4,j}^U|z_{3,j}^{Ap}, z_{2,j}^{Ap}, z_{1,j}^{Ap}, J = j, t, Z^{Ad}) \times Pr(z_{5,j}^U|z_{4,j}^U, z_{3,j}^{Ap}, z_{2,j}^{Ap}, z_{1,j}^{Ap}, J = j, t, Z^{Ad})$. The terms in the right hand side of above equation can be evaluated based on an intruder's knowledge of the unknown variables. Such knowledge can include estimated probability distributions, model estimates, and details about the SDL procedures used to generate the data set for release. Incorporating these details is discussed in Section 4.

3.2.3 $Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C|z_j^C, J = j, t, Z^{Ad})$

The expression $Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C|z_j^C, J = j, t, Z^{Ad})$ corresponds to the probability of observing values on variables in C in every record but the j^{th}, given the values observed for variables in C on the j^{th} record, given the j^{th} record belongs to the target, the target's information, and values in Z^{Ad}. Assuming independence between records, $Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C|z_j^C, J = j, t, Z^{Ad})$ can be expressed as the product of probabilities $Pr(z_i^C|z_j^C, J = j, t, Z^{Ad})$ over records $i = 1, \cdots, j-1, j+1, \cdots, n$. Notice that the idenpendence assumption implies $Pr(z_i^C|z_j^C, J = j, t, Z^{Ad}) = Pr(z_i^C|J = j, t, Z^{Ad})$. If this product is multiplied and divided by $Pr(z_j^C|z_j^C, J = j, t, Z^{Ad})$, and this term is then substituted into the expression for $Pr(J = j|t, Z)$, simplifications lead to this being equivalent to substituting $1/Pr(z_j^C|z_j^{Ad})$ into the original expression (Reiter 2005).Using this simplification, we procede by developing an expression for $Pr(z_j^C|z_j^{Ad})$ based on the properties of variables in $C = (Ap, U)$. Details are presented in Section 4.

4. Component Formulation for Three Intruders

As noted above, various assumptions about an intruder's knowledge and behavior can be incorporated into assessing disclosure risk under the Duncan and Lambert, Reiter framework. This gives the flexibility to consider various scenarios of intruder knowledge and behavior. As in Reiter (2005), we characterize intruder knowledge and behavior as naive, average, or SDL. The naive intruder is one who only possesses posterior information, or information available from the released data. The SDL intruder is one who has accurate and fairly detailed knowledge of the statistical disclosure limitation (SDL) method used. From the agency's point of view, the naive intruder might represent a best case scenario and the SDL intruder a worst case scenario. Between these two extremes lies an average intruder who has a combination of knowledge about the data prior to its release and the SDL method used, but neither completely. The developments in this research include assessing the disclosure risk associated with the SDL intruder when the SDL method is to produce synthetic data using predictions from model estimates. In this section, details for assessing disclosure risk associated with an SDL intruder when the SDL method is to generate synthetic data using quantile regression predictions, hot deck imputation, and rank swapping. Details for assessing disclosure risk associated with a naive and an average intruder can be found in Reiter (2005).

Consider data set Y to contain variables $Y_1, ..., Y_7, Y^{Ad}$, where Y^{Ad} are the variables copied directly into the synthetic data set, Y^{Ap} are the available perturbed variables, and $Y^U = Y^{Up}, Y^{Up}$ are the unavailable variables that do not get perturbed and do get perturbed. Using our proposed SDL procedure, we produce Z for release. Specifically,

1. variables in Z^{Ad} remain unperturbed (possibly re-categorized), i.e., $Z^{Ad} = Y^{Ad} \Rightarrow t^{Ad} = Z_t^{Ad}$, where Z_t^{Ad} is the target's record in the released data,

2. values for z^{Ap}_1 are generated conditional on Z^{Ad} using quantile regression predictions at randomly selected quantiles, i.e., $Q_{z^{Ap}_{1j}}(\tau^*|Z^{Ad}) = Z^{Ad}_j \beta_1(\tau^*_{1j})\epsilon_1$,

3. values for z^{Ap}_2 are generated conditional on y^{Ap}_1, z^{Ap}_1, and Z^{Ad} using quantile regression predictions at randomly selected quantiles, i.e., $Q_{z^{Ap}_{2j}}(\tau^*|Z^{Ad}, y^{Ap}_1) = \left(Z^{Ad}_j, z^{Ap}_{1j}\right)' \beta_2(\tau^*_{2j}) + \epsilon_2$,

4. values for z^{Ap}_3 and z^{Ap}_4 are generated using hot deck imputation and rank swapping, matching based on Mahalanobis distance between synthetic and original values of Z_1 and Z_2

5. values for z^{Up}_5 are generated conditional on y^{Ap}_1, z^{Ap}_1, y^{Ap}_2, z^{Ap}_2, and Z^{Ad} using quantile regression predictions at randomly selected quantiles, i.e., $Q_{z^{Ap}_{5j}}(\tau^*|Z^{Ad}, y^{Ap}_1, y^{Ap}_{2j}) = \left(Z^{Ad}_j, z^{Ap}_{1j}, z^{Ap}_2\right)' \beta_5(\tau^*_{5j}) + \epsilon_5$,

6. values for z^{Up}_6 are generated using hot deck imputation and rank swapping, and

7. and values for z^{Ud}_7 are left unperturbed in the released data.

Assuming an intruder has the information and knows some details about the QR and HD+RS procedures used, the probability of identification can be estimated accordingly. Here we discuss the $Pr(J = j|t, Z^{Ad})$ component. In the remainder of this section, we discuss components A, B, and C.

Component $Pr(J = j|t, Z^{Ad})$ can be estimated using information in t and Z^{Ad}. This information is available prior to data release and values are not perturbed in the released data. This component is estimated using $Pr(J = j|t, Z^{Ad}) = \frac{1}{n_t}$ when $z^{Ad}_j = t^{Ad}$ and zero otherwise, where n_t is the number of records released in Z with $z^{Ad}_j = t^{Ad}$ (Reiter 2005). This formulation of $Pr(J = j|t, Z^{Ad})$ can be used for any intruder type or SDL method used.

4.1 Formulation of $C_{SDL} = Pr(z^{Ap}|J = j, t, Z^{Ad})$

Under the SDL scheme outlined above, we assume $z^{Ap} = (z^{Ap}_1, z^{Ap}_2, z^{Ap}_3, z^{Ap}_4)$ as the available perturbed variables where z^{Ap}_1 and z^{Ap}_2 are quantile regression predictions and z^{Ap}_3 and z^{Ap}_4) are hot deck imputations with rank swapping.

The *SDL* intruder is assumed to know details about the quantile regression models. Specifically, that $z^{Ap}_1 = \hat{y}^{Ap}_1 = Z^{Ad}\hat{\beta}_{1,\tau_1}$, but not the exact values of τ_1, and $z^{Ap}_2 = \hat{y}^{Ap}_2 = (Z^{Ad}\ z^{Ap}_1)^T\hat{\beta}_{2,\tau_2}$, but not the exact values of τ_2. Similary, the intruder is assumed to know that each of z^{Ap}_3 and z^{Ap}_4 are values from the actual data set selected based on hot deck imputation dependent on distances between (y^{Ap}_1, y^{Ap}_2) and (z^{Ap}_1, z^{Ap}_2) and swapped based on ranks within some distance δ of the value identified by hot deck. The iintruder is assumed not to know the distance δ.

Using this information one can formulate C_{SDL} as $C_{SDL} = Pr(z^{Ap}_j|J = j, t, Z^{Ad}) = Pr(z_{1j}, z_{2j}, z_{3j}, z_{4j}|J = j, t, Z^{Ad}) = Pr(z_{1j}|J = j, t, Z^{Ad}) \times Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad}) \times Pr(z_{3j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad}) \times Pr(z_{4j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad})$, for every j^{th} released record $j = 1, ..., r$. Note that in each of the four terms, the conditionaing variables include those used to generate the corresponding synthetic value. By considering the details of the SDL procedures, each term in C_{SDL} can be developed, as follows.

4.1.1 C1: z^{Ap}_1 and z^{Ap}_2

The variables z^{Ap}_1 and z^{Ap}_2 are available and perturbed using predictions from quantile regression models. An expression for C1 is developed as if the intruder knows the values of τ, then extended to incorporate the more likely case that the intruder does not know these values. Results in Koenker (2002) indicate that regression parameter estimates at quantile τ are asymptotically Normal, with mean and variance dependent on quantile τ. Assuming independent and identically distributed errors, $\sqrt{n}(\hat{\beta}_n(\tau) - \beta(\tau))\tilde{\rightarrow}N(0, \omega^2(\tau))$, where $\omega^2(\tau) = \frac{\tau(1-\tau)}{f^2(F^{-1}(\tau))}$. For practical purposes, we assume $\hat{\beta}_n \sim N\left(\beta(\tau), \frac{\omega^2(\tau)}{n}\right)$, which implies $X\hat{\beta}_n \sim N\left(X\beta(\tau), \frac{\omega^2(\tau)}{n}X'X\right)$. Details are presented in Koenker (2002).

If synthetic values are the predictions $z_{1j} = \hat{y}_{1j,\tau_{1j}} = z^{Ad}_j\hat{\beta}_1(\tau_1)$, and we assume the intruder knows the value of τ_{1j} for every $j = 1, ..., r$, then s/he can formulate $Pr(z_{1j}|J = j, t, Z^{Ad})$ for each record $j = 1, ..., r$ to be:

$$Pr(z_{1j}|J = j, t, Z^{Ad}) = \phi_{1j,\tau_{1j}} = \phi\left((z_{1j} - t_{1j})/(\omega(\tau_{1j})\sqrt{z^{Ad^T}_j z^{Ad}_j/n})\right). \tag{3}$$

Supposing the intruder does not know values of τ_1, this value can be estimated. The estimates could be set to a single constant, could be randomly selected, or otherwise estimated according to values in the released data. The intruder's estimate, $\hat{\tau}_{1,intruder}$, can be substituted into Equation 3 for τ_{1j} to obtain an estimate of $Pr(z_{1j}|J = j, t, Z^{Ad})$.

Values for z_2 are generated similarly, hence the formulation of $Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad})$ follows. The quantile regression model is $z_{2j} = \hat{y}_{2j,\tau_2} = \left(Z^{Ad}z_{1j}\right)'\hat{\beta}_2(\tau_2)$. Thus by substituting the intruder's estiamte of $\hat{\tau}_{2,intruder}$, into Equation 4, an estimate of $Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad})$ can be obtained:

$$Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad}) = \phi_{2j,\tau_{2j}} = \phi\left((z_{2j} - t_2)/(\omega(\tau_{2j})\sqrt{(Z_j^{Ad}\ z_{1j})(Z_j^{Ad}\ z_{1j})^T/n})\right) \tag{4}$$

if $t^{Ad} = z_j^{Ad}$ and zero otherwise.

4.1.2 Alternative to C1: z_1^{Ap} and z_2^{Ap}

In Equations 3 and 4 we are willing to assume the approximate asymptotic normality of quantile regression parameter estimates. This is generally acceptable in applications with large data sets, since the regression estimates are based on over 10,000 and up to millions of records. If a data base contains a small number of records or if too many records contain all zeros or very small values, the assumptions may be unreasonable.

Alternatively, the target's t_1 and t_2 predicted values, $\hat{t}_1$ and $\hat{t}_2$, could be computed using quantile regression estimates from the released data set and the target values of t^{Ad}. The intruder could then compare the target's predicted values $\hat{t}_1$ and $\hat{t}_2$ to values z_1 and z_2 released in Z. Among records with equal available and unperturbed variable values, $t^{Ad} = Z^{Ad}$, it would be reasonable to consider identifying the target with any record containing z_1 and z_2 values within some range of $\hat{t}_1$ and $\hat{t}_2$. If the intruder is willing to consider any records within an amount $\gamma_1 > 0$, say, of $\hat{t}_1$, then all records within this distance to $\hat{t}_1$ have equal probability of belonging to the target. If the intruder simultaneously considers only records with values of z_2 within $\gamma_2 > 0$ of $\hat{t}_2$, then only the records with values z_1 and z_2 within the intervals $(\hat{t}_1 \pm \gamma_1)$ and $(\hat{t}_2 \pm \gamma_2)$ are considered as potential matches with the target. Records with z_1 and z_2 values within these intervals have equal probability of belonging to the target. Suppose there are n_{t_1,t_2} such records, then the joint probability of observing z_1^{Ap} and z_2^{Ap} conditional on the j^{th} record belonging to the target, the information in t, and the values in Z^{Ad} can be formulated as

$$Pr(z_{1j}, z_{2j}|J = j, t, Z^{Ad}) = Pr(z_{1j}|J = j, t, Z^{Ad})Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad}) = 1/n_{t_1,t_2} \tag{5}$$

if $t^{Ad} = Z^{Ad},\ z_1 = \hat{t}_1 \pm \gamma_1, z_2 = \hat{t}_2 \pm \gamma_2$ and zero otherwise.

4.1.3 C2: z_3^{Ap} **and** *z_4^{Ap}*

The variables z_3^{Ap} and z_4^{Ap} are available to the intruder and perturbed using hot deck imputation with rank swapping. We can formulate $Pr(z_{3j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad})$ and $Pr(z_{4j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad})$ based on details of the SDL procedures. In the hot deck procedure for our example data set, we identify matching records in the original data set based on the Mahalanobis distance between values $(Y^{Ad}, Y_1^{Ap}, Y_2^{Ap})$ and $(Z^{Ad}, Z_1^{Ap}, Z_2^{Ap})$. For synthetic record j, among any original records with $y^{Ad} = z^{Ad}$ the distance $d_(i,j) = d\left[\begin{pmatrix} y_{1i}^{Ap} \\ y_{2i} \end{pmatrix}, \begin{pmatrix} z_{1j} \\ z_{2j} \end{pmatrix}\right]$ is computed for each original record. If $d_(i,j)$ is the smallest for original record i, the sample ranks r_{3i} and r_{4i} are computed for y_{3i} and y_{4i}, respectively. Ranks r_{3i}^* and r_{4i}^* are drawn from a discrete Uniform distributions over the intervals $(r_{3i} - \delta_3, r_{3i} + \delta_3)$ and $(r_{4i} - \delta_4, r_{4i} + \delta_4)$ and values with ranks r_{3i}^* and r_{4i}^*.

This leads us to formulate the probability of observing z_3 and z_4 given $z_{2j}, z_{1j}, J = j, t,$ and Z^{Ad} independent of the distance between the target values and the values in Z. The argument above implies that the probability of t_3 and t_4 being imputed into the released data set can be based on the rank swapping portion of the procedure alone. Recall, the probability $Pr(z_{3j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad})$ is conditional on the j^{th} record belonging to the target. If we assume that the j^{th} record belongs to the target, then we can assume the values t_3, t_4, were swapped with values y_3, y_4 having ranks that were randomly selected from a Uniform distribution over the intervals $(r_{t_3} - \delta_3, r_{t_3} + \delta_3)$ and $(r_{t_4} - \delta_4, r_{t_4} + \delta_4)$, respectively. Therefore, for observations z_{3j} and z_{4j} to have been imputed, the ranks in the original record r_{3i} and r_{4i} must fall in the intervals, $(r_{t_3} - \delta_3, r_{t_3} + \delta_3)$ and $(r_{t_4} - \delta_4, r_{t_4} + \delta_4)$, respectively. If we assume values in Z_3 and Z_4 have approximately the same ranks as values in Y_3 and Y_4, then the ranks of values z_{3j} and z_{4j}, r_{3j}^* and r_{4j}^*, are also in that interval. We can formulate the conditional probability of observing the z_{3j} and z_{4j} to be equal to the probability of selecting their ranks r_{3j}^* and r_{4j}^* from the intervals $(r_{t_3} - \delta_3, r_{t_3} + \delta_3)$ and $(r_{t_4} - \delta_4, r_{t_4} + \delta_4)$ and 0 for ranks not in these intervals. This can be written $Pr(z_{3j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad}) = 1/(2\delta_3)$ if $t^{Ad} = z_j^{Ad}$ and $r_{3j}^* \in (r_{t_3} \pm \delta_3)$ and zero otherwise. Similarly, $Pr(z_{4j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad}) = 1/(2\delta_4)$ if $t^{Ad} = z_j^{Ad}, r_{4j}^* \in (r_{t_4} \pm \delta_4)$ and zero otherwise. Since rank swapping is done independently to obtain values z_3 and z_4, we can simply multiply the terms to obtain the joint probability $Pr(z_{3j}, z_{4j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad})$.

Combining the components for z_{1j}, z_{2j}, z_{3j}, and z_{4j} we arrive at the following expression for C_{SDL}: $Pr(z^{Ap}|J = j, t, Z^{Ad}) = \phi_{1j,\tau_{1j}}\phi_{2j,\tau_{2j}}\frac{1}{\delta_3}\frac{1}{\delta_4}$ if $t^{Ad} = Z^{Ad}, r_{3j}^* \in (r_{t_3} \pm \delta_3), r_{4j}^* \in (r_{t_4} \pm \delta_4$ and zero otherwise. We can use the ideas presented above to formulate components corresponding to additional available perturbed variables, Z^{Ap}, when conditional quantile regression predictions and hot deck imputation with rank swapping are used to generate values in synthetic records. When other SDL methods are used to generate synthetic data, it seems feasible to extend the ideas presented here and in Reiter (2005) to formulate components of C_{SDL}. In particular, it should be straight forward to extend the formulation of $Pr(z_{1j}|J = j, t, Z^{Ad})$ and $Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad})$ in Equations 3 and 4 to synthetic values that are generated using predictions from any conditional model, provided distibutional properties of model estimates and subsequent predictions are known or can be derived. Reiter (2005) presents possible formulations of $Pr(z_j^{Ap}|J = j, t, Z^{Ad})$ when swapping, re-categorizing, and noise addition are used.

4.2 B: $B_{SDL} = Pr(z_j^U | z_j^{Ap}, J = j, t, Z^{Ad})$

Variables z_j^U are variables that are unavailable before release. They include z_{5j}^{Up} which is perturbed using quantile regression predictions before release, z_{6j}^{Up} which is perturbed using hot deck and rank swapping before release, and z_{7j}^{Ud} w hich is unperturbed before release. Recall, an SDL intruder is assumed to know how values are generated for each variable in Z. As for C_{SDL}, we assume the intruder uses this knowledge to formulate B_{SDL}. This implies the intruder's joint conditional probability of observing z_5, z_6, z_7, will be formulated using information about the statistical disclosure limitation procedure. In our hypothetical data set, Z^U is comprised of z_5, z_6, z_7, the variables with values unknown to the intruder prior to data release. We rewrite the joint distribution of z_5, z_6, z_7, as a series of conditional distributions. For every $j = 1, ..., r$, $B_{SDL} = Pr(z_j^U | z_j^{Ap}, J = j, t, Z^{Ad})$ can be written as $Pr(z_{5j}, z_{6j}, z_{7j} | z_j^{Ap}, J = j, t, Z^{Ad}) =$

$$Pr(z_{5j} | z_{1j}, z_{2j}, J = j, t, Z^{Ad}) \times Pr(z_{6j} | z_{1j}, z_{2j}, J = j, t, Z^{Ad}) \times Pr(z_{7j} | J = j, t, Z^{Ad}). \tag{6}$$

In the proposed SDL procedure, z_5 is generated using quantile regression predictions conditional on z_{1j}, z_{2j}, and Z^{Ad}, z_6 is generated using hot deck and rank swapping conditional on z_{1j} and z_{2j}, and z_7 is left unperturbed in the released data. Based on this, we divide the variables in U into perturbed and unperturbed just as for available variables, i.e. $z_5, z_6 \in Up$ and $z_7 \in Ud$. In the following paragraphs, we first consider $Pr(z_{5j} | z_j^{Ap}, J = j, t, Z^{Ad})$, then $Pr(z_{6j} | z_j^{Ap}, J = j, t, Z^{Ad})$, and finally $Pr(z_{7j} | J = j, t, Z^{Ad})$.

4.2.1 B1: z_{5j}^{Up}

Variable z_{5j}^{Up} is unavailable to the intruder before release and is perturbed using quantile regression predictions before release. We consider formulating the conditional probability of observing z_5 in a similar fashion as the conditional probability corresponding to z_1. Unlike the probability of observing z_1, the intruder does not have information on z_5 (or any variables in U) prior to data release, i.e. s/he does not have values t_5, t_6, or t_7. Therefore, $Pr(z_{5j} | z_j^{Ap}, J = j, t, Z^{Ad}) = \phi_{5j,\tau_{5j}}$ cannot be evaluated as in Equation 3 using z_{5j} and t_5.

The intruder does know that z_{5j} is generated using $z_{5j} = \left(Z_j^{Ad}, z_{1j}, z_{2j}\right)' \hat{\beta}_5(\tau_{5j})$, for every $j = 1, ..., r$, but does not know the value $\hat{\beta}_5(\tau_{5j})$. This parameter estimate could be estimated by the intruder by fitting the corresponding quatile regression model using values in Z. The estimate of $\hat{\beta}_5(\tau_{5j})$ could then be used to compute a predicted value for the target on this variable, $\hat{t}_5$, based on values of t_1, t_2, and t^{Ad}.

The SDL intruder's predicted $\hat{t}_5$ is not the exact value released by the agency on the target's record (due to estimated quantile regression estimates). How similar or different these values are will depend on the accuracy of the intruder's estimate and the value $\hat{\beta}_5(\tau_{5j})$ estimated by the agency using the original data set. In other words, if the relationship between z_5 and z_1, z_2, and Z^{Ad} are preserved very accurately in the released data, this would result in accurate estimated coefficients for the intruder, and an accurate predicted value of the target's released value. The intruder can either act as if $\hat{t}_5$ is equal to the target's released value or account for additional error introduced by estimating the model coefficients using Z rather than Y. We consider the former scenario, but recognize resulting probability estimates may differ when considering the latter.

Using the estimated value of $\hat{t}_5$, the intruder can choose to act as if this estimate is the target's value of t_5 and procede as above for z_1 and z_2. Plugging $\hat{t}_5$ in for t_5 to evaluate the Normal density, we obtain the following expression for the probability of observing z_{5j} for every $j = 1, ..., r$:

$$Pr(z_{5j} | z_{2j}, z_{1j}, J = j, t, Z^{Ad}) = \phi_{5j,\tau_{5j}} = \phi\left((z_{5j} - t_5)/(\omega(\tau_{5j})\sqrt{(Z_j^{Ad}\ z_{1j}\ z_{2j})(Z_j^{Ad}\ z_{1j}\ z_{2j})^T/n})\right) \tag{7}$$

when $t^{Ad} = z_j^{Ad}$ and zero otherwise, where $\tau_{5j} = \tau_{5j}^*$ if the intruder knows the value of randomly drawn τ_{5j}^* used to generate z_{5j} and it equals $\hat{\tau}_{5j,intruder}^*$ if the intruder estimates the value of randomly drawn τ_{5j}^* used to generate z_{5j}.

Alternatively, the intruder may compare the target values of t^{Ad} and $\hat{t}_5$ to values in Z. Suppose there are n_{t_5} records with $t^{Ad} = Z^{Ad}$ and $z_5 = \hat{t}_5 \pm \gamma_5$, some $\gamma_5 > 0$, then the probability can be formulated as $Pr(z_{5j} | z_j^{Ap}, J = j, t, Z^{Ad}) = 1/n_{t_5}$ when $t^{Ad} = Z^{Ad}$ and $z_5 = \hat{t}_5 \pm \gamma_5$ and zero otherwise.

The size of γ_k depends on the amount of error the intruder attributes to estimating the regression coefficient estimates. The intruder would likely be more willing to act more certain about $\hat{t}_5$ as an estimate of t_5 if the model parameters are well estimated using the released data set. However, the intruder might want to make computation reasonable and choose γ_k so that n_{t_5} is not too large or too small. Namely, if there are few observations close to the predicted target value, then a value of γ_k that is somewhat large would ensure n_{t_5} is not too small. In particular, one would not want to take the chance of eliminating potential matches that could be the target through a choice of n_{t_5} that is too small.

4.2.2 B2: z_{6j}^{Up}

Variable z_{6j}^{Up} is unavailable to the intruder before release and is perturbed using hot deck and rank swapping before release. Recall, hot deck and rank swapping are combined to generate values for z_6 in the released data. We consider formulating $Pr(z_{6j} | z_{1j}, z_{2j}, J = j, t, Z^{Ad})$ in a similar manner as the corresponding probability statements for z_3 and z_4. In the case of z_6, however, we do not have the target value t_6 to use to compute the rank of this variable in the target's record. To use the previous formulation, the intruder would need to estimate the rank of t_6. This could be done by estimating the value of t_6, according to some model, then computing its rank relative to values of z_6 in the relased data, or perhaps by modeling the ranks themselves, r^*_{6j}, conditional on other variables in Z. Investigating the best way to estimate the rank of t_6 in the target's record is left to outside research. In a simulation study, we consider estimating t_6 based on a model and computing the rank of the estimated value with respect to values of z_6.

Regardless of how this is done, if the intruder obtains an estimate of the rank of t_6, $\hat{r}_{t_6}$ say, then s/he can use this value to evaluate $Pr(z_{6j}|z_{1j}, z_{2j}, J = j, t, Z^{Ad}) = 1/(2\delta_6)$ when $t^{Ad} = Z^{Ad}, r^*_{6j} \in (\hat{r}_{t_6} \pm \delta_6)$ for some $\delta_6 > 0$ and zero otherwise.

4.2.3 B3: z_{7j}

Variable z_{7j} is unavailable to the intruder before release and is unperturbed before release. To compute $Pr(z_{7j}|J = j, t, Z^{Ad})$, we rely on an argument presented in Reiter (2005). The author presents the conditional probability as an integral of the joint probability of z^U_j and y^U_j over values of y^U_j as follows:

$$Pr(z^U_j|J = j, t, Z^{Ad}) = \int Pr(z^U_j|y^U_j, J = j, t, Z^{Ad})Pr(y^U_j|J = j, t, Z^{Ad})dy^U_j.$$

The author points out that if variables in U remain unperturbed, i.e. $U = Ud$, then $Pr(z^U_j|y^U_j, J = j, t, Z^{Ad}) = 1$, so the entire integral integrates to 1. For our purposes, we set $Pr(z_{7j}|J = j, t, Z^{Ad}) = 1$, for all $j = 1, ..., r$, assuming the intruder knows values on z_7 are all left unperturbed from their original values.

4.3 A: $A_{SDL} = Pr(z^C_1, ..., z^C_{j-1}, z^C_{j+1}, ..., z^C_r|z^C_j, J = j, t, Z^{Ad})$

Variables z^C_i are variables the intruder cannot know whit certainty after release. Variables in z^{Ap}_i and z^U_i are in z^C_i, so these variables include $z^{Ap}_{1i}, z^{Ap}_{2i}, z^{Ap}_{3i}, z^{Ap}_{4i}, z^{Up}_{5i}, z^{Up}_{6i}$, and z^{Ud}_{7i}. This component computes the probability associated with all records except the target's. In Section 2.2.2, we introduce A_{SDL} as the third term in the right hand side of Equation 2, where $A_{SDL} = Pr(z^C_1, ..., z^C_{j-1}, z^C_{j+1}, ..., z^C_r|z^C_j, J = j, t, Z^{Ad})$.

Assuming records are independent, this expression simplifies to the product of conditional probabilities (Reiter 2005):

$$Pr(z^C_1, ..., z^C_{j-1}, z^C_{j+1}, ..., z^C_r|z^C_j, J = j, t, Z^{Ad}) = \prod_{\substack{i=1,\ldots,r \\ i \neq j}} Pr(z^C_i|z^C_j, J = j, t, Z^{Ad}) \tag{8}$$

Since records are independent, then for $i \neq j$, $Pr(z^C_i|z^C_j, J = j, t, Z^{Ad}) = Pr(z^C_i|z^{Ad}_i)$. Substituting this into Equation 8 and rewriting the product, we obtain $Pr(z^C_1, ..., z^C_{j-1}, z^C_{j+1}, ..., z^C_r|z^C_j, J = j, t, Z^{Ad}) = \prod_{i=1,...,r} Pr(z^C_i|z^{Ad}_i)/Pr(z^C_j|z^{Ad}_j)$. Substituting this into Equation 1 for $Pr(z^C_1, ..., z^C_{j-1}, z^C_{j+1}, ..., z^C_r|z^C_j, J = j, t, Z^{Ad})$ results in further simplifications that occur from summing over all records in the denominator of 1. As a result, the above substitution is equivalent to substituting $1/Pr(z^C_j|z^{Ad}_j)$ into 1 for $Pr(z^C_1, ..., z^C_{j-1}, z^C_{j+1}, ..., z^C_r|z^C_j, J = j, t, Z^{Ad})$.

This implies that only $Pr(z^C_j|z^{Ad}_j)$ is needed to compute A_{SDL}. This probability is decomposed into conditional probabilities according to available perturbed, unavailable perturbed, and unavailable unperturbed variables as in the preceding sections: $Pr(z^C_j|z^{Ad}_j) = Pr(z^{Ap}_j, z^{Up}_j, z^{Ud}_j|z^{Ad}_j) = Pr(z^{Ap}_j|z^{Ad}_j)Pr(z^{Up}_j|z^{Ap}_j, z^{Ad}_j)Pr(z^{Ud}_j|z^{Ap}_j, z^{Up}_j, z^{Ad}_j)$. Recall that for unavailable, unperturbed variables z^{Ud}_j, the corresponding probability is set to 1, resulting in further simplification of $Pr(z^C_j|z^{Ad}_j)$ to

$$Pr(z^C_j|z^{Ad}_j) = Pr(z^{Ap}_j|z^{Ad}_j)Pr(z^{Up}_j|z^{Ap}_j, z^{Ad}_j). \tag{9}$$

Under the SDL method of this section, the probability becomes $Pr(z^C_j|z^{Ad}_j = Pr(z^{Ap}_{1j}|z^{Ad}_j)\ Pr(z^{Ap}_{2j}|z^{Ap}_{1j}, z^{Ad}_j) \times Pr(z^{Ap}_{3j}, z^{Ap}_{4j}|z^{Ap}_{2j}, z^{Ap}_{1j}, z^{Ad}_j)\ Pr(z^{Up}_{5j}|z^{Ap}_j, z^{Ad}_j)\ Pr(z^{Up}_{6j}|z^{Up}_{5j}, z^{Ap}_j, z^{Ad}_j)$.

To evaluate this probability, consider the same SDL methods as those used as in the previous sections. The probabilities in Equation 9 are no longer conditioned on target information, t, or $J = j$. Therefore, instead of using values from the target information t, values in each record are used. For $j = 1, ..., r$ the probabilities are listed here, followed by a brief discussion:

$$\begin{array}{lrll}
z^{Ap}_j: & Pr(z_{1j}|z^{Ad}_j) = & \tilde{\phi}_{1j,\tau_{1j}} & = \phi\left((z_{1j} - \hat{z}_{1j,\tau})/(\omega(\tau_{2j})\sqrt{Z^{Ad^T}_j Z^{Ad}_j/n})\right) \\
 & Pr(z_{2j}|z_{1j}, z^{Ad}_j) = & \tilde{\phi}_{2j,\tau_{2j}} & = \phi\left((z_{2j} - \hat{z}_{2j,\tau_{2j}})/(\omega(\tau_{2j})\sqrt{(Z^{Ad}_j\ z_{1j})(Z^{Ad}_j\ z_{1j})^T/n})\right) \\
 & Pr(z_{3j}|z_{2j}, z_{1j}, z^{Ad}_j) = & 1/(2\delta_3), & \hat{r}_{3j} \in (r_{3j} - \delta_3, r_{3j} + \delta_3) \\
 & Pr(z_{4j}|z_{2j}, z_{1j}, z^{Ad}_j) = & 1/(2\delta_4), & \hat{r}_{4j} \in (r_{4j} - \delta_4, r_{4j} + \delta_4) \\
 & & & \\
z^{Up}_j: & Pr(z_{5j}|z_{2j}, z_{1j}, z^{Ad}_j) = & \tilde{\phi}_{5j,\tau_{5j}} & = \tilde{\phi}\left((z_{5j} - \hat{z}_{5j,\tau_{5j}})/(\omega(\tau_{5j})\sqrt{(Z^{Ad}_j\ z_{1j}\ z_{2j})(Z^{Ad}_j\ z_{1j}\ z_{2j})^T/n})\right) \\
 & Pr(z_{6j}|z_{2j}, z_{1j}, z^{Ad}_j) = & 1/(2\delta_6), & \hat{r}_{6j} \in (r_{6j} - \delta_6, r_{6j} + \delta_6)
\end{array} \tag{10}$$

The formulations of the components of A_{SDL} are quite similar to the components for B_{SDL} and C_{SDL} since the SDL method and intruder's knowledge are the same. They differ in due to conditioning only on observed values in record j, rather than on the target's information and on $J = j$, the j^{th} record belonging to the target. The terms $\hat{z}_{kj,\tau_{kj}}$ are defined as before, as the quantile regression predictions for variable k at the τ_{kj} quantile, computed using released values and estimated parameter estimates obtained from the released data. The probabilities associated with the variables that imputed using hot deck and rank swapping remain at $1/(2\delta_k)$ when the rank of $\hat{z_{kj}}$, $\hat{r}_{kj}$, falls in the interval $(r_{kj} - \delta_k, r_{kj} + \delta_k)$. The value of δ_k is not likely released by the agency. The intruder trades off taking large values of δ_k to cover the true match and a small value of δ_k that is more computationally feasible. Future work could examine ways to estimate δ_k.

5. Summary

Under the framework for computing disclosure risk presented in Duncan and Lambert (1986, 1989) and Reiter (2005), components of disclosure risk were formulated based on various levels on intruder knowledge and decisions. A summary is presented in this section. The methods have been implemented in an application to a Public Use Microdata Sample from the U.S. Census Bureau. Simulation and case study results will be reported elsewhere.

The framework for measuring disclosure assumes that an intruder computes the probability of identifying a target in the released data set, which is expressed as $Pr(J = j|t, Z^{Ad})$. Disclosure risk is equated with the probability of identification. If the agency can control the probability of identification to be low for a target, then the disclosure risk is also low. Alternatively, if the probability of identication is the same accross a large number of records, this may prevent the intruder from identifying any record as the target's, resulting in low disclosure risk as well.

To assess disclosure risk, we consider the various types of information or knowledge an intruder has before data release, information gained after release, and various decisions the intruder can make about how to calculate the probability of identification. Such decisions are based on the level of information s/he possesses. Disclosure risk is divided into extreme cases based on an *SDL* intruder and a *naive* intruder. An *average* intruder is also considered to give insight into a possibly more common type of intruder. By computing disclosure risk for each type of intruders, we hope to cover a wide range of possibilities, enabling the agency to evaluate risk in a worst case scenario, a best case scenario, and a more common scenario. We have also included something like a best best case scenario using the *super naive* intruder, who bases the probability of identification only on the number of records with matching available unperturbed variables.

Disclosure risk can also be computed for the intruder that makes decisions to compute the components in a simpler manner than s/he has information for. For example, an intruder with accurate and detailed information about the SDL method used can choose to compute C_{SDL}, B_{SDL}, and A_{SDL} to obtain the probability of identification. Alternatively, an SDL intruder can choose to compute C_{SDL}, but use B_{avg} and A_{avg}, or even set these components to 1, to compute the probability of identification. The average intruder has options too. S/he can compute all components using the average formulations (C_{avg}, B_{avg}, and A_{avg}), or can compute any of these components at the naive level or set any of them equal to 1. The naive intruder can only choose to compute C_{naive}, B_{naive}, and A_{naive} or set these components equal to 1. The naive intruder, however, cannot choose to increase the amount of prior knowledge used to compute any component. In total, there are 3^3 options of combinations of A, B, and C available to the SDL intruder, or 4^3 options if we include setting any component to 1. There are 2^3 (or 3^3) options for the average intruder, and one option (or 2^3) for the naive intruder.

Intruders with other levels of knowledge exist and can make choices to formulate the probability of identification in a different way than we have. We hope to account for the best and worst case scenarios based on SDL knowledge, average knowledge, and naive knowledge by using the formulations presented in this section.

Acknowledgments

This work was supported in part by a Census Dissertation Fellowship (Census Task Order YA132307SE0304). This article is the responsibility of the authors and does not necessarily reflect the opinion of the U.S. Census Bureau or any of its employees. Special thanks go to Sam Hawala, Laura Zayatz, and Tommy Wright.

REFERENCES

Duncan, G. T. and Lambert, D. (1986). Disclosure-Limited Data Dissemination. *Journal of the American Statistical Association.* **86 393** 10-18.

Duncan, G. T. and Lambert, D. (1989). The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistics.* **7 2** 207-217.

Huckett, J.C., and Larsen, M.D. (2007). Microdata simulation for confidentiality protection using regression quantiles and hot deck. *Proceedings of the Survey Research Methods Section, ASA.*

Huckett, J.C., and Larsen, M.D. (2008). Combining Methods to Create Synthetic Microdata: Quantile Regression, Hot Deck, and Rank Swapping. *Proceedings of Statistical Society of Canada.*

Koenker, R. (2002). *Quantile Regression.* Econometric Society Monograph Series, Cambridge University Press. New York, New York.

Reiter, J.P. (2005) Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100, 1103-1112