VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 123–146, http://dx.doi.org/10.2478/jos-2014-0007

Disclosure-Protected Inference with Linked Microdata Using a Remote Analysis Server

James O. Chipperfield[1]

Large amounts of microdata are collected by data custodians in the form of censuses and administrative records. Often, data custodians will collect different information on the same individual. Many important questions can be answered by linking microdata collected by different data custodians. For this reason, there is very strong demand from analysts, within government, business, and universities, for linked microdata. However, many data custodians are legally obliged to ensure the risk of disclosing information about a person or organisation is acceptably low. Different authors have considered the problem of how to facilitate reliable statistical inference from analysis of linked microdata while ensuring that the risk of disclosure is acceptably low. This article considers the problem from the perspective of an Integrating Authority that, by definition, is trusted to link the microdata and to facilitate analysts' access to the linked microdata via a remote server, which allows analysts to fit models and view the statistical output without being able to observe the underlying linked microdata. One disclosure risk that must be managed by an Integrating Authority is that one data custodian may use the microdata it supplied to the Integrating Authority and statistical output released from the remote server to disclose information about a person or organisation that was supplied by the other data custodian. This article considers analysis of only binary variables. The utility and disclosure risk of the proposed method are investigated both in a simulation and using a real example. This article shows that some popular protections against disclosure (dropping records, rounding regression coefficients or imposing restrictions on model selection) can be ineffective in the above setting.

Key words: Confidentiality; remote analysis; record linkage; statistical disclosure control.

1. Introduction

Large amounts of microdata are collected by data custodians in the form of censuses and administrative sources. Often, data custodians will collect different information on the same individual. Many important questions can be answered by linking microdata collected by different data custodians. For this reason, there is very strong demand from analysts, within government, business and universities, for linked microdata. However, many data custodians are legally obliged to ensure the risk of disclosing information about a person or organisation is acceptably low. For simplicity, in the rest of this article it is assumed that there are only two data custodians and the *linked microdata* are the result of linking two sets of microdata collected by the two data custodians. Potential analysts of

[1] Senior Research Fellow, National Institute for Applied Statistics Research Australia, University of Wollongong, and Assistant Director, Methodology Division, Australian Bureau of Statistics, Canberra, ACT, 2617, Australia. Email: james.chipperfield@abs.gov.au

the linked microdata are the two data custodians and noncustodians (e.g., academics, members of the public). There are two reasons the disclosure risks are significantly greater if an analyst of the linked microdata is also a data custodian. First, because data custodians commonly collect name and address, any additional information that can be inferred about a record on the microdata it collected, can be directly associated with the person who provided it. Second, a data custodian can use information on the linked microdata collected to disclose information about a person or organisation on the linked microdata that was collected by the other data custodian.

There has been some work in the literature on managing the disclosure risks from analysts who are also data custodians. When unique identifiers, such as name and address, are available, record linkage techniques (see Herzog et al. 2007) are frequently used to identify records belonging to the same individual. Secure Record Linkage (SRL) (see, for example, Churches and Christen 2004) suggests a way in which a third party can link microdata without each data custodian disclosing the identity of nonlinked records to the other data custodian and without the data custodians revealing any sensitive information to the third party. The data custodians attach a unique record identifier to their microdata (e.g., random number) and agree on a common way of encrypting the linking variables, which are sent to the third party to perform the record linkage. The third party links the microdata and returns the record identifiers of the linked pairs to the data custodians. Therefore, each data custodian could identify the names and addresses of the people who were linked, which in turn could disclose sensitive information (e.g., knowing a person's record has been linked to an unemployment register discloses the person is unemployed). For many data custodians, such as the Australian Bureau of Statistics (ABS), revealing such information would be a breach of their legal obligations and would mean that SRL is not a viable option. If instead the third party was allowed access to linking variables (e.g., name and address), the linkage could be of much higher quality, since clearly unencrypted linking variables are more useful in identifying matches than encrypted linking variables. It would be interesting to study the extent to which encryption of linking variables reduces the quality of the linkage.

Once the linked pairs are determined, each data custodian will need to ensure that any statistical output from the linked microdata has an acceptable disclosure risk. Secure computation algorithms allow data custodians to compute matrix operations, such as those involved in regression, from linked microdata without sharing individual records (see, for example, Karr et al. 2009). Among the major limitations of this approach are that it relies on SRL, allows only datacustodians to analyse the microdata (i.e., non-data custodians cannot perform analysis) and that it is currently limited to a certain set of models. Alternatively, Kohnen and Reiter (2009) consider the novel problem of how data custodians, without sharing sensitive variables, can together produce synthetic linked microdata for public use. Limitations of this approach are that synthetic data can be time consuming to produce and that it can be hard to guarantee that the synthetic data do not distort some important relationships.

In contrast to the above approaches discussed in the literature, this article considers a more practical and straightforward approach to managing disclosure risk from linked microdata. In particular, this article considers the presence of a so-called Integrating Authority (IA) that is trusted to perform the following roles:

1. Link microdata collected by two data custodians.
2. Maximise the inherent utility or value of the linked microdata. This may include application of consistent standards and classifications, statistical editing and imputation.
3. Allow analysts to access the linked microdata in order to fit models.
4. Ensure the level of disclosure risk of the regression output is acceptable to the data custodians.

The IA is allowed to observe the microdata collected by the data custodians. The data custodians do not mask the microdata they provide to the IA in any way. The data custodians not only have access to the microdata they provided to the IA but, as analysts, they also have access to the regression output released by the Integrating Authority.

There are at least three benefits to an IA. First, the IA manages the complexity involved in linking microdata and managing disclosure risk – this is important since many data custodians do not have the specialised capability in, for example, standardising linking fields, editing and imputation, record linkage and data access. Second, since the IA observes the linking fields, it is possible to conduct a clerical review on the set of links and to refine the method of record linkage. This essential task appears impractical when linking fields are encrypted. Third, a more optimal trade-off between disclosure risk and the utility of the analysis is possible. With an IA, only the disclosure risk of the regression output needs to be managed; under the alternatives mentioned above, the disclosure risk must be managed from record linkage to construction of the regression output itself.

There are some major potential disadvantages of the IA framework. First, some data custodians may be prohibited by law, from disclosing information to any another organisation. This would mean the IA framework would not apply. Second, fulfilling the role of an IA is potentially a costly exercise. This may lead the IA to pass this cost burden onto analysts by charging a substantial fee for access. Moreover, it is the IA that decides how to fulfil its roles in any given situation. For example, the IA decides which variables to include on the linked microdata and how analysts will access the linked microdata (e.g., public use file or via a remote server, as discussed below). These decisions may suit some analysts but not others.

Once the record linkage step is completed by the IA, its next step is to facilitate access to the microdata. In this article, the IA releases regression output via a remote analysis server (see Reiter 2002, Gomatam et al. 2005, Sparks et al. 2008, Lucero and Zayatz 2010). A simple model for a remote server is:

1. An analyst submits a query, via the Internet, to the analysis server.
2. The analysis server processes the analyst's query on the linked microdata. The statistical output (e.g., regression coefficients) is modified or restricted in order to ensure the risk of disclosure is acceptably low.
3. The analysis server sends the modified output, via the Internet, to the analyst.

One key protection against disclosure afforded by remote analysis is that the analyst is restricted from viewing the microdata. However, an analyst may attempt to use the regression output to infer the value of variables on the linked microdata. Such attempts are commonly called *data attacks*. Once the value of these variables is inferred, the attacking

analyst can attempt one of the well-understood methods of disclosure (e.g., attribute disclosure through linkage); for a review see Shlomo (2007). The IA can provide analysts with disincentives to conducting attacks in the first place. For example, analysts could be required to sign confidentiality agreements to access to the remote server. If the agreement is violated by an analyst, access to the server could be revoked.

This article is about how an IA can manage the risk of a data custodian successfully attacking the linked microdata. A data custodian's attack would involve using the microdata it supplied to the IA and the regression output released by the remote server to infer the value of variables about a person or organisation that were collected by the other data custodian. Data custodians will commonly collect name and address, which means if such an attack is successful, the value of any variables that are inferred could be directly attributed to the person or organisation who provided that information. In other words, disclosure occurs automatically after a successful attack.

The problem of managing the disclosure risk of regression output released via a remote server has been the subject of significant recent attention in the literature. The literature on this problem focuses on the situation where there is a single data custodian responsible for managing access to its microdata (i.e., unlinked microdata). In the case of remote analysis for model fitting, most effort has been directed at linear regression. Gomatam et al. (2005) considered imposing restrictions to stop analysts reconstructing coefficients for a sensitive linear model, an example of which is a model with highly accurate predictions of a sensitive characteristic (see Bleninger et al. 2010 for an empirical investigation). Taking a completely opposite approach, Dwork and Smith (2009) describe the concept of *differential privacy*, which imposes no restrictions but instead relies on perturbation of statistical output alone to manage the disclosure risk. Many authors have considered imposing both restrictions and perturbation (e.g., Sparks et al. 2008); this article takes such an approach. One limitation of a remote server is that analysts are restricted to using the set of statistical analysis procedures that are supported by the remote server. This article only briefly mentions the more moderate disclosure risk of attacks made by noncustodians since, as mentioned, there is a considerable literature on this problem.

Section 2 describes how a data custodian may attack the linked microdata when the remote server naively releases standard regression output for models that are fitted to binary data. Section 3 proposes simple protections that an IA can implement in a remote server to reduce the success rate of these attacks. Section 4 evaluates the utility and disclosure risk of the proposed approach in a real situation and in a simulation. Section 5 makes some final comments.

2. Attacks Without Any Protection

This section describes how a data custodian can attack the linked microdata if the remote server naively releases standard regression output. Consider an IA linking microdata collected by two data custodians, referred to as A and T. Data Custodian A is the *attacker* and Data Custodian T is the *target*.

This article makes the assumption that all links between records are correct (i.e., each pair of records that are linked correspond to the same person or organisation) and that the name and address of all linked records are known to Data Custodian A. In practice, linkage

is rarely perfect and it is well known that errors arising during the linkage process provide some level of protection against disclosure (see Ch. 18 in Herzog et al. 2007). From the perspective of managing disclosure risk, the assumption that linkage errors do not arise is conservative.

Many authors distinguish between variables that are sensitive (e.g., income) and those that are not sensitive, where only the risk of disclosing sensitive variables needs to be managed. However, the legislation that guides how the Australian Bureau of Statistics, and many other data custodians, manages disclosure risk does not distinguish between sensitive and nonsensitive variables.

Let $\mathcal{D}$ be a set of records from the linked microdata comprising n records: a binary outcome variable y and a vector of K binary covariates $\mathbf{x}$. For the ith record, define $(y_i, \mathbf{x}_i)$ where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki}, \ldots, x_{Ki})'$ and $i = 1, \ldots, n$. Data Custodian T collected y and the K_T column vector $\mathbf{x}_T$ and Data Custodian A collected the K_A column vector $\mathbf{x}_A$ so that $\mathbf{x} = (\mathbf{x} = (\mathbf{x}_T', \mathbf{x}_A')')'$ and $K = K_T + K_A$. In other words, if we define $\mathbf{X} = (\mathbf{X}_T, \mathbf{X}_A) = (\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots \mathbf{x}_n)'$, Data Custodian T supplied $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_n)'$ and the $n \times K_T$ matrix $\mathbf{X}_T$ and Data Custodian A collected the $n \times K_A$ matrix $\mathbf{X}_A$. Therefore we may now write $\mathcal{D} = (\mathbf{y}, \mathbf{X})$.

An attack by Data Custodian A involves using regression output released by the remote server and $\mathbf{X}_A$ to infer the value of one or more elements of $(\mathbf{y}, \mathbf{X}_T)$. Therefore, for the purposes of this article, if a variable on the linked microdata is collected by both data custodians (e.g., a linking variable), it is defined as a covariate in $\mathbf{x}_A$.

In general, a good strategy for Data Custodian A's attack on a record involves ensuring $\mathbf{x}_A$, used in the calculation of the statistical output, uniquely identifies the target record on the linked microdata. This ensures there is 1–1 mapping between the target record's value of $\mathbf{x}_A$ and name and address. As Data Custodian A collected $\mathbf{X}_A$, this could readily be achieved.

Noncustodians present much less of a disclosure risk. Firstly, since they do not have access to $\mathbf{X}_A$, they can only use the statistical output released by the remote server in an attack. Secondly, even if an attack was able to reconstruct $(y_j, \mathbf{x}_j)$, attributing the jth record to a person or organisation is more difficult without name and address (see Skinner and Shlomo 2008).

Subsections 2.1, 2.2 and 2.3 describe attacks using standard regression output, including estimates of regression coefficients, estimates of their variance and test statistics, respectively.

2.1. Regression Coefficients

The standard estimate of the regression coefficient β for models fitted to binary variables (e.g., logistic regression, linear regression), denoted by $\hat{\beta}$, is obtained by solving the score equation

$$Sc(\beta; \mathcal{D}) = 0, \tag{1}$$

where $Sc(\beta) = \Sigma_i \mathbf{x}_i'(y_i - \mu_i)$ and $\mu_i = g(\mathbf{x}_i'\beta)$ for some link function g. It is well known that fitting a model to $\mathcal{D}$ is equivalent to fitting a model to the C counts contained in the vector $\mathbf{n}$, where $\mathbf{n} = \{n_c : c = 1, \ldots, C\}$ and n_c is the number of records belonging to the

cth pattern in $(y, \mathbf{x})$ (see McCullagh and Nelder 1989). As an aside, if y was instead a multinomial response with M categories, the appropriate score function would involve $M - 1$ equations of the form of (1). A multinomial response model fits into the framework developed here, but for simplicity we do not consider it further.

This section shows how Data Custodian A can attack $(\mathbf{y}, \mathbf{X}_T)$ – this involves using $\hat{\beta}$ and $\mathbf{X}_A$ in an attempt to infer the value of one or more elements of $(\mathbf{y}, \mathbf{X}_T)$.

2.1.1. Solving the Estimating Equations from a Single Model

Consider Data Custodian A substituting $\hat{\beta}$ into (1) and then attempting to solve for some elements of $(\mathbf{y}, \mathbf{X}_T)$. If the number of patterns in $\mathbf{x}_A$ is C_A, Data Custodian A's attack can exploit the following:

1. The K constraints imposed by $\hat{\beta}$ through (1)
2. Knowledge of $\mathbf{X}_A$
3. $(\mathbf{y}, \mathbf{X}_T)$ has only binary elements.

This attack can be as simple as conducting a grid search. Other more sophisticated search techniques could also be used. Of course this search could be more targeted if, for instance, Data Custodian T were to release frequency counts of y or $\mathbf{x}_T$ to the public. For example, the ABS, as potential Data Custodian T, releases frequency counts from its Census microdata after the counts have been perturbed by a small amount.

2.1.2. Solving Estimating Equations from Multiple Models

This attack involves fitting different models to the same set of data values in $\mathcal{D}$ (i.e., the same set of records and variables) by:

1. Changing the dependent variable
2. Changing the link function (e.g., linear, logistic and probit)
3. Transforming variables (e.g., creating an interaction term).

The regression coefficients for each fitted model impose additional constraints on $(\mathbf{y}, \mathbf{X}_T)$ via (1). The idea behind this attack is to impose sufficient constraints so that Data Custodian A can solve for one or more elements of $(\mathbf{y}, \mathbf{X}_T)$.

Example 1: Solving for all unknowns. Denote the data values in $\mathcal{D}$ by $\mathbf{Z} = (\mathbf{X}, \mathbf{y}) = (\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_n)'$, where z_{im} to be the mth element of $\mathbf{z}_i$. Consider Data Custodian A fitting the mth model where the outcome variable for the ith record is $y_i^{(m)} = z_{im}$ and the covariate for the ith record is $\mathbf{x}_i^{(m)}$, which is obtained by dropping z_{im} from $\mathbf{z}_i$. Denote the standard estimate of the regression coefficients from the mth model by $\hat{\beta}^{(m)}$ and denote $\hat{\mu}_i^{(m)} = g(\mathbf{x}_i^{(m)'}\hat{\beta}^{(m)})$. Data Custodian A's attack involves solving for one or more elements of $(\mathbf{y}, \mathbf{X}_T)$ given $\mathbf{X}_A$, $\hat{\beta}^{(m)}$ and the constraint

$$\Sigma_i \mathbf{x}_i^{(m)}(y_i^{(m)} - \hat{\mu}_i^{(m)}) = 0, \tag{2}$$

for $m = 1, \ldots, M$. Clearly, as M increases so does the number of constraints.

Example 2: Solving for unknowns in one estimating equation. Continuing *Example 1*, consider the lth estimating equation in (2) when Data Custodian A fits M models such that $y_i^{(m)} = y_i$ and the lth element of $\mathbf{x}_i^{(m)}$ is by definition x_{il}, for all $m = 1, \ldots, M$.

Further consider that Data Custodian A collected x_l so that it knows which $H = \Sigma_i x_{il}$ records contribute to the lth estimating equation. The constraint imposed by the lth estimating equation in (2) reduces to

$$\Sigma_{i,x_{il}=1} y_i - \Sigma_{i,x_{il}=1} \hat{\mu}_i^{(m)} = 0,$$

for $m = 1, \ldots, M$. This imposes M constraints on the $H \times (K_T + 1)$ unknowns for the H records contributing to the lth estimating equation. This number of unknowns could be considerably less than *Example 1*. An extreme example is when $H = 1$, which means there are $(K_T + 1)$ unknowns and M constraints.

Example 3: Imposing more constraints by creating a new variable. If Data Custodian A collected the variable t, it could repeat the attack in *Example 1* or *2* but where y_i is replaced with $y_{new,i} = y_i t_i$ for all i. By imposing the additional constraint $y_{new,i} = 0$ if $t_i = 0$, Data Custodian A can focus on solving y_i for *only* records with $t_i = 1$. This additional constraint could considerably reduce the number of unknowns.

2.1.3. Counts

Consider if Data Custodian A regresses $\mathbf{y}$ on $\mathbf{x} = \mathbf{x}_A$ and aims to infer $\mathbf{T} = \Sigma_i \mathbf{x}_i' y_i$, which are counts of y in the margins of $\mathbf{x}$. Given $\hat{\beta}$ and (1), this is straightforward since $\mathbf{T} = \Sigma_i \mathbf{x}_i' \mu_i$. The disclosure risks of frequency counts are well known (see, for example, Shlomo 2007). Counts of one would lead to disclosure. Counts of one can also be obtained through differencing, as discussed below.

2.1.4. Differencing

A standard differencing (see, for example, Gomatam et al. 2005; Shlomo 2007) attack involves fitting the same model to two sets of records that are identical except that one record is dropped from one of the sets. Data Custodian A can be sure only the target record is dropped if the dropping condition uniquely identifies the record and if it collected all the variables in the dropping condition. Differences in the estimated regression coefficients from the two models can be used in an attempt to infer the values of the dropped record's variables.

Example 4: Differencing attack by dropping a record. Consider if Data Custodian A wants to infer y_r, the value of y for rth record. Data Custodian A can fit a linear regression model with $\mathbf{x} = \mathbf{x}_A$ before and after dropping the rth record. Denote the value of the estimated regression coefficients before and after dropping the rth record by β_o and $\beta_{o(r)}$, respectively. Also denote $\mathbf{y}_{(r)}$ and $\mathbf{X}_{(r)}$ by $\mathbf{y}$ and $\mathbf{X}$ after removing the rth row, respectively. Since Data Custodian A knows β_o, $\beta_{o(r)}$, $\mathbf{X}_{(r)}$ and $\mathbf{X}$, it can calculate $\mathbf{S}_o = \mathbf{X}'\mathbf{X}\beta_o = \mathbf{X}'\mathbf{y}$ and $\mathbf{S}_{o(r)} = \mathbf{X}_{(r)}'\mathbf{X}_{(r)}\beta_{o(r)} = \mathbf{X}_{(r)}'\mathbf{y}_{(r)}$ and take the difference $\mathbf{S}_{o(r)} - \mathbf{S}_o = \mathbf{x}_r' y_r$. Since y_r is the only unknown, Data Custodian A can infer it directly.

2.1.5. Fishing

Fishing attacks involve fitting two models that are only different in one small way. Of interest is whether the two sets of coefficients are the same or whether they are different; how the coefficients change is not of interest. An example is given below.

Example 5: Fishing by slightly changing the definition of a variable. Consider linked microdata where Data Custodian A collected a variable for small area geography and age in single years and Data Custodian T collected a sensitive variable. Data Custodian A would know if there was one record in a particular small area with age equal to 100 years and may seek to infer the value of the sensitive characteristic for the record. Data Custodian A could fit two models to the records in the small area which are exactly the same, except that the first includes a binary covariate that takes the value one when age is between 40 and 100 *and* the sensitive characteristic is present and the second model includes a binary covariate that takes the value one when age is between 40 and 99 *and* the sensitive characteristic is present. If the regression coefficients from these two models are different, Data Custodian A infers that the 100-year-old has the condition; otherwise Data Custodian A infers that the 100-year-old does not have the condition.

2.2. *Estimated Variance of Regression Coefficients*

The estimated variance of $\hat{\beta}$ is $\widehat{Var}(\hat{\beta}; \mathcal{D}) = (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}$, where $\hat{\mathbf{V}}$ is diagonal with ith element $v^{-1}\left(\partial\mu/\partial\eta\right)^2$ evaluated at $\mathbf{x} = \mathbf{x}_i$ and $\beta = \hat{\beta}$, v is the variance function for the model, and $\eta = \mathbf{x}'\beta$. Given $\hat{\beta}$, $\widehat{Var}(\hat{\beta}; \mathcal{D})$ can impose up to $K(K-1)/2$ constraints on $\mathbf{X}$. These constraints could be exploited to assist with an attack on estimated regression coefficients. Consider the simple linear regression model where $\widehat{Var}(\hat{\beta}; \mathcal{D}) = \hat{\phi}(\mathbf{X}'\mathbf{X})^{-1}$ which, after taking the inverse and multiplying by released dispersion parameter $\hat{\phi}$, gives the table of counts $\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_T'\mathbf{X}_T & \mathbf{X}_T'\mathbf{X}_A \\ \mathbf{X}_A'\mathbf{X}_T & \mathbf{X}_A'\mathbf{X}_A \end{pmatrix}$. Many of the attacks in Subsection 2.1 (e.g., differencing attacks and fishing) can be used against $\widehat{Var}(\hat{\beta}; \mathcal{D})$. They are not discussed further here.

2.3. *Other Statistical Output*

Regression analysis would normally include exploratory data analysis, use of test statistics and graphical plots to assess the model fit. Univariate and multivariate exploratory analysis involving binary variables will often involve frequency counts, which are well known to be a disclosure risk (see references below). Such work goes beyond the scope of this article, but will form the subject of future work.

Statistics used to assess model fit or goodness-of-fit (see Hosmer and Lemeshow 2000), say $t = t(\hat{\beta}, \mathcal{D})$, are functions of the microdata $\mathcal{D}$ and an estimate of β. Again, many of the attacks in Subsection 2.1 (e.g., differencing attacks and fishing) can be used against t. They are not discussed further here.

Graphical diagnostics are frequently used to assess model fit. The disclosure risk of plotting record-level values is high and has been considered by many authors (see O'Keefe and Good 2009 and O'Keefe et al. 2012). Consider if a remote server releases $\hat{\beta}$ and a plot which shows that the predicted value for a record is p. Given $\mathbf{x}$ has only binary elements, there will in general be only a single value of $\mathbf{x}$ such that $p = \mu(\mathbf{x})$. Furthermore, if the record has a unique value for $\mathbf{x}_A$ on the linked microdata, then Data Custodian A can infer $\mathbf{x}_T$ for the person about which the record relates.

3. Attacks in the Presence of Protections

This section proposes some simple protections against the attacks described in the previous section. The objective of these protections is to significantly reduce the likelihood of a successful attack while making a small impact on the utility of the analysis. Subsections 3.1 and 3.2 consider protections by imposing a general set of restrictions and by introducing uncertainty into regression coefficients, respectively. Subsection 3.3 considers attacks on estimated regression parameters in the presence of these protections. Subsections 3.4 and 3.5 describe protections for the variance of the estimated regression parameters and for diagnostic test statistics, respectively.

3.1. Protection: Imposing General Restrictions

Several restrictions are suggested below. These restrictions do not necessarily defend against a particular attack, but are designed to significantly hinder attacks while resulting in only a minor reduction in utility. When designing a set of restrictions to manage disclosure risk, it quickly becomes clear that a series of legitimate regression models *could* be indistinguishable from a sophisticated data attack. Therein lies the challenge: not restricting the former while thwarting the latter. This challenge is discussed in detail by Cox et al. (2011).

Some analysts may have good reasons for fitting a model which is not permitted by the set of restrictions below. The IA could decide to relax some restrictions if: the analyst promises to fit a small number of predefined models (this could be verified by the IA); if the IA believes errors, such as incorrect or missed links, in the linked microdata provide substantial protection; or if the analysis has high utility. For obvious reasons, the IA would be more willing to relax restrictions for analysts who are not data custodians, as long as they promise not to share the regression output publically.

If the IA is not willing to relax one or more restrictions so that an analyst may fit a particular model, the IA may provide the analyst with an alternative mode of access to the linked microdata. One example would be for the IA to provide the analyst with the C counts required to fit the model, though the counts will almost certainly need to be carefully perturbed to manage the risk of disclosure.

Some possible general restrictions include:

(a) Limit the number of model covariates, K, by imposing the restriction that $K < 30$. Models with a large number of covariates may impose considerable constraints on unknowns. In very few cases would legitimate analysis be impacted by this restriction.
(b) Impose a minimum number of observations or covariate patterns by imposing the restrictions $n \geq 50$ and $C > 50$. This restriction aims to ensure a minimum number of unknowns. Remembering that C is the number of counts to which the model is fitted, $C \leq 2^K$ effectively means that $K > 5$.
(c) Adjusted $R - squared < 0.95$ (see also Gomatam et al. 2005). Other cut-off values can be considered. Inferential disclosure occurs when a model's prediction of a sensitive variable, y, is highly accurate and all covariates for the target record are known (e.g., $\mathbf{x} = \mathbf{x}_A$). This restriction is designed to prevent inferential disclosure. This protection will rarely be required since accurate predictions of binary outcomes

are rare. (Aside: inferential disclosure is fundamentally based on model assumptions. Some would argue that inferences which rely on model assumptions cannot lead to disclosure, because there is uncertainty about whether the model assumptions are true.)

(d) Each variable in the model must be non-zero for at least ten records. As all variables are binary this means $\Sigma_i x_{ik} \geq 10$ and $\Sigma_i(1 - x_{ik}) \geq 10$ for all k, $\Sigma_i y_i \geq 10$ and $\Sigma_i(1 - y_i) \geq 10$. This provides some protection against attacking a single estimating Equation (see *Example 2*) by ensuring there will be a minimum of ten unknowns.

(e) $(C - C_A) \geq 10K$. This ensures that there are ten times the number of unknown counts than there are constraints imposed by the estimating equation.

(f) New variables may only be created by multiplying two variables originally on the microdata as long as both variables were collected by the same data custodian. This restriction aims to prevent a data custodian from, almost arbitrarily, reducing the number of unknowns as in *Example 3*.

(g) Exclude variables from the linked file if they have limited analytic value. This limits the potential prior knowledge a data custodian can use in attacks. This decision must be made by the IA after consultation with potential analysts.

(h) Restrict variables which are naturally only useful as model covariates (e.g., marital status, age, sex, geography) from being dependent variables. This will hamper attempts to solve the estimating equation by changing the choice of dependent variable (see point 1 in Subsection 2.1.2). See also Gomatam et al. 2005 for another justification for this restriction.

It makes sense to impose data custodian-specific restrictions (e.g., see (e) above) because the disclosure risk naturally depends upon which data custodian is performing the attack. For data custodian-specific restrictions to make sense it must be assumed that there is restricted (e.g., to publications) sharing of regression coefficients between data custodians and that data custodians are aware of what regression coefficients they are able to share. What if this assumption is not realistic? The implication is that if one data custodian is restricted from fitting a model then all data custodians and non-data custodians must be restricted from fitting the model. In other words – *restriction for one means restriction for all*.

While the details are not within the scope of this article (for details see O'Keefe and Chipperfield 2013), the IA will need to decide what restrictions, if any, to place on subsetting records (i.e., defining the records in $\mathcal{D}$). If there is no restriction on subsetting, a data custodian may be able to arbitrarily target records to drop in differencing attacks. On the other hand, the flexibility of subsetting is very important since it allows analysts to make inferences about a specific population of interest.

If the number of models that are fitted is allowed to be arbitrarily high, the corresponding set of constraints may be such that an attacking data custodian can *solve the estimating equation*. Therefore it is worth mentioning a basic indicator of the risk of this attack succeeding. Consider when Data Custodian A fits its mth model to $C_{(m)}$ counts, where $C_{(m)}$ is the same as C but for the mth model and $C_{A(m)}$ is the same as C_A but for the mth model. Consider $L_A = \Sigma_m L_{A(m)}$, where $L_{A(m)} = C_{(m)}^{-1} C_{A(m)}$. The numerator of $L_{A(m)}$ is the number of constraints Data Custodian A can impose on the $C_{(m)}$ counts (see point 2 in

Subsection 2.1.1) to which the mth regression model was fitted. When $L_A > 1$ there are potentially more constraints than unknown counts, at which point the IA could perhaps audit the models fitted by Data Custodian A. Refining this indicator and developing similar indicators for other attacks would be an interesting line of future work.

3.2. Protection: Introducing Uncertainty into the Released Regression Coefficients

Two simple ways of introducing uncertainty into regression coefficients are now mentioned. The first protection is that a different random sample of records is dropped (see Sparks et al. 2008) for every distinct model that is fitted. Specifically, for each $k = 1, \ldots, K$, one record with $x_k = 1$ is randomly selected and dropped from $\mathcal{D}$. This means K records will be dropped. Denote $\mathcal{D}_{drop}$ to be $\mathcal{D}$ after dropping records in this way. Estimates of regression coefficients will not be biased by dropping records in this way, since it does not affect the distribution of y conditional on $\mathbf{x}$. As many applications involving linked microdata have a large number of records, dropping records in this way will generally only have a small impact on the accuracy of estimates. Note that dropping a completely random sample of records for every model fitted (see Sparks et al. 2008) provides limited protection in the present setting. Consider dropping 50 randomly selected records as a protection against the attack in *Example 2*, where $n = 50{,}000$ and $H = 5$ so that $x_k = 1$ for only five records. Since it is unlikely that $x_k = 1$ for any of the dropped records, it is equally unlikely that the attack in *Example 2* will be affected by dropping records in this way.

The second protection involves adding noise (for a review see O'Keefe and Chipperfield 2013) to the RHS of (1). Consider the estimator $\hat{\beta}^*$ of β, obtained by solving

$$Sc(\beta; \mathcal{D}_{drop}) = \mathbf{E}^*, \tag{3}$$

where the microdata used in the regression are $\mathcal{D}_{drop}$ not $\mathcal{D}$, $\mathbf{E}^* = (E_1^*, \ldots, E_k^*, \ldots, E_K^*)'$, $\mathbf{E}_k^* = \phi u_k^*$, ϕ is a scaling factor for the perturbation that needs to be set by the integrating authority and u_k^* s are independently generated variables from the uniform distribution on the range $(-1,1)$. Other distributions can be considered. The regression coefficients $\hat{\beta}$ are perturbed via $\mathbf{E}^*$. The value for ϕ is best determined through empirical investigation and simulation, which is discussed below. The distribution for u_k^* is bounded so that the impact of perturbation is bounded. The contribution of a record to the kth estimating equation is in the range $(-1,1)$, which is also the range of the perturbation, u_k^*. As many attacks attempt to uncover the values of variables for a single record, this is arguably a minimum degree of perturbation.

The distribution of the perturbations in $\mathbf{E}^*$ are independent so that $Var(\mathbf{E}^*)$ is a diagonal matrix. The joint distribution of $\mathbf{E}^*$ across different models should also be independent with an important exception: the same values of $\mathbf{E}^*$ should be used if exactly the same model is fitted. This condition stops estimation of $\hat{\beta}$ by fitting exactly the same model a number of times and averaging over the $\hat{\beta}^*$s.

3.3. Attacks Using the Released Estimated Regression Coefficients

Here we revisit the attacks of Section 2 in the presence of the protections mentioned above. It is assumed here that ϕ and the rules for dropping records are in the public domain.

3.3.1. Solving the Estimating Equation

Define $\hat{\beta}^{(m)*}$ to be the same as $\hat{\beta}^{(m)}$ except that it is obtained by solving (3) rather than (1). Consider solving the estimating equation in *Example 1* but where the regression parameter, $\hat{\beta}^{(m)*}$ instead of $\hat{\beta}^{(m)}$, is released. Define $\mathcal{D}^{(m)}_{drop}$ to be $\mathcal{D}$ after randomly dropping records for the mth model. Data Custodian A's attack now involves finding, over all possible subsets $\mathcal{D}^{(m)}_{drop}$ of $\mathcal{D}$, a unique solution for one or more elements of $\mathbf{y}$ given

$$-\phi 1 \leq \left\{ \Sigma_{i \in D^{(m)}_{drop}} \mathbf{x}^{(m)}_i (y^{(m)}_i - \hat{\mu}^{*(m)}_i) \right\} \leq \phi 1, \tag{4}$$

for $m = 1, \ldots, M$, where $\hat{\mu}^{*(m)}_i = g(\mathbf{x}^{(m)'}_i \hat{\beta}^{*(m)})$ and 1 is a K vector of 1s.

The protection provided by perturbation and dropping records depends upon the many possibly interacting factors implicit in (4). This makes it difficult to make any general conclusions about the protections they provide against disclosure. Clearly, the protection provided by perturbation is driven by ϕ. When looking at (4), it is clear that as ϕ increases the interval becomes wider and the probability of a unique solution (i.e., disclosure) becomes smaller. The method of dropping records would ideally prevent strict constraints being imposed on the terms in (4). If $y = 1$ for 99% of records, then an attack could assume, with high probability of being correct, that $y = 1$ for all dropped records. Making this assumption would impose a further constraint on the unknown values of y – in particular, if the first element of $\mathbf{x}$ was a constant, then the first element of $\mathbf{A} = \Sigma_{i \in D^{(m)}_{drop}} \mathbf{x}^{(m)}_i y^{(m)}_i$ in (4) would be constant over m. The first element of $\mathbf{A}$ could no longer be assumed to be constant if there was some uncertainty about how many records were dropped (e.g., instead of dropping one randomly selected record with $x_k = 1$, drop $1, 2, \ldots,$ or T randomly selected records with $x_k = 1$ with probability $1/T$).

3.3.2. Counts

Consider how $\hat{\beta}^*$ protects against estimating $\mathbf{T} = \Sigma_{i \in \mathcal{D}} \mathbf{x}'_i y_i$. If Data Custodian A regresses $\mathbf{y}$ on $\mathbf{x} = \mathbf{x}_A$, it can compute $\hat{\mathbf{T}}^* = \Sigma_{i \in \mathcal{D}} \mathbf{x}_i \hat{\mu}^*_i$, where $\hat{\mu}^*_i = g(\mathbf{x}'_i \hat{\beta}^*)$. Data Custodian A knows the minimum and maximum value for the counts in $\mathbf{T}$ are given by the corresponding elements of $\mathbf{T}_{min} = \hat{\mathbf{T}}^* - (\phi + K)\mathbf{1}$ and $\mathbf{T}_{max} = \hat{\mathbf{T}}^* + \phi \mathbf{1}$, respectively. The '$K$' in the expression for $\mathbf{T}_{min}$ reflects the fact that Data Custodian A knows that up to K records could be dropped from each estimating equation.

3.3.3. Differencing

Consider how perturbation protects against differencing attacks on counts (see *Example 4*), assuming for the moment that no records are randomly dropped (i.e., $\mathcal{D}_{drop} = \mathcal{D}$). Consider if Data Custodian A regresses $\mathbf{y}$ on $\mathbf{x} = \mathbf{x}_A$ before and after dropping the rth record. Accordingly define $\mathcal{D}_{(r)}$, $\mathbf{T}_{(r)} = \Sigma_{i \in \mathcal{D}_{(r)}} \mathbf{x}'_i y_i$, $\hat{\mathbf{T}}^*_{(r)} = \Sigma_{i \in \mathcal{D}_{(r)}} \mathbf{x}_i \hat{\mu}^*_{i(r)}$, where $\hat{\mu}^*_{i(r)} = g(\mathbf{x}'_i \hat{\beta}^*_{(r)})$, and $\hat{\beta}^*_{(r)}$ to be exactly the same as $\mathcal{D}$, T, $\hat{\mathbf{T}}^*$ and $\hat{\beta}^*$, respectively, except that they are computed after the rth record is dropped. Data Custodian A can compute an estimate of $\mathbf{x}'_r y_r$ by

$$\Delta_{(r)} = \hat{\mathbf{T}}^* - \hat{\mathbf{T}}^*_{(r)}. \tag{5}$$

If any element of $\Delta_{(r)}$ has magnitude greater than 2ϕ, Data Custodian A can infer that $y_r = 1$. It is also not hard to see that if $y_r = 0$ this differencing attack will never succeed. This means the success rate of this attack depends upon the probability that $y = 1$ for the target records. It is also not hard to see that, as K increases and ϕ decrease, the probability of this attack succeeding increases.

Now consider the same differencing attack but where records are randomly dropped, as discussed previously. Denote $\mathcal{D}_{(r)drop}$ to be the result of randomly dropping records from $\mathcal{D}_{(r)}$. Now $\hat{\beta}^*$ and $\hat{\beta}^*_{(r)}$ are calculated from $\mathcal{D}_{drop}$ and $\mathcal{D}_{(r)drop}$ instead of $\mathcal{D}$ and $\mathcal{D}_{(r)}$, respectively. Of course, the IA does not reveal which records are dropped so that $\mathcal{D}_{(r)drop}$ and $\mathcal{D}_{drop}$ are not known to Data Custodian A. Accounting for this uncertainty, it is easy to show if any element of $\Delta_{(r)}$ has magnitude greater than $2\phi + K$ (the difference between $\mathbf{T}_{min}$ and $\mathbf{T}_{max}$), Data Custodian A can infer that $y_r = 1$.

3.3.4. Fishing

Randomly dropping records as described above provides an effective protection against fishing attacks since, for every distinct model that is fitted, a different random sample of records is dropped. This will mean, continuing with *Example 5*, that the regression coefficients for the two models will be different whether or not the 100-year-old has the condition. Only if the same model is fitted repeatedly (i.e., the chosen link function, the set of records, and dependent and independent variables are all the same) should the same set of records be dropped. Otherwise this protection can be removed by averaging.

3.4. Variance of Estimated Regression Coefficients

Given the perturbation and model distributions are independent, the sandwich estimator for the variance of $\hat{\beta}^*$ is

$$\widehat{Var}\left(\hat{\beta}^*; \mathcal{D}_{drop}\right) = \widehat{Var}\left(\hat{\beta}; \mathcal{D}_{drop}\right) + (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}_{drop} Var_*(\mathbf{E}^*)(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}_{drop}, \tag{6}$$

The first term in (6) is the estimated variance of the standard estimator of β obtained from solving (1), but based on $\mathcal{D}_{drop}$ rather than $\mathcal{D}$. An analytic expression for the first term is $(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}_{drop}$, where $(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})_{drop}$ is $\mathbf{X}'\hat{\mathbf{V}}\mathbf{X}$ but based on $\mathcal{D}_{drop}$ rather than $\mathcal{D}$. Alternatively the first term can be calculated from $\mathcal{D}_{drop}$ using the Bootstrap or Jackknife (see Chambers and Skinner 2003 p. 105). The second term in (6) measures the variation due to perturbation where it is easy to show, using the variance of the Uniform distribution, that $Var_*(\mathbf{E}^*)$ is diagonal with kth element $var_*(\phi u^*_k) = \phi^2/3$. The analyst can make valid inferences about β using $\hat{\beta}^*$ and (6), without knowing anything about the perturbation itself. It is interesting to note that the first and second terms of (6) are $O(n^{-1})$ and $O(n^{-2})$ respectively, which means that the impact of perturbation on variance is small.

Using the same reasoning as in Subsection 2.2, releasing (6), where the first term is computed analytically, would represent a high risk of disclosure. Instead consider computing the first term using the Jackknife. Denote θ as the analytic variance estimate of $\hat{\beta}$ and denote $\hat{\theta}$ as the corresponding Jackknife variance estimate of θ. The Jackknife estimate has a level of uncertainty due to the process, denoted by v, of allocating selection units to replicate groups. In particular, the coefficient of variation of $\hat{\theta}$ due to this process is $CV_v\left(\hat{\theta}\right) \approx 2(R-1)^{-1}$, where R is the number of replicate groups, $CV_v(\hat{\theta}) = Var(\hat{\theta})\hat{\theta}^{-2}$

(see Shao and Tu 1995, p. 196) and m/n is negligible. As long as R is not too large, this uncertainty in $\hat{\theta}$ will mask θ. This means that computing the first term in (6) using the Jackknife will mask the entire RHS of (6). For example, if the Jackknife standard error estimate is 0.2 and is based on $R = 50$, a 95% confidence interval for the estimate is $(0.31, 0.46)$.

It is difficult to see how (6) could be used in a differencing attack or be used to impose any constraint that would be useful to help solve the estimating equation. Since (6) is based on $\mathcal{D}_{drop}$ it is protected from fishing attacks. A further protection is to release only the diagonal elements of (6) so that only the variances of the regression coefficients are released.

3.5. *Other Statistical Output*

Given $\hat{\beta}^*$ instead of $\hat{\beta}$ is released, it makes sense that an analyst would be interested in $t^* = t(\hat{\beta}^*, \mathcal{D}_{drop})$ rather than t. The statistic t^* for the adjusted R^2, leverage, dispersion parameter and the Hosmer Lemeshow and chi-squared statistics will have their usual interpretation (i.e., replacing $\hat{\beta}$ and $\mathcal{D}$ with $\hat{\beta}^*$ and $\mathcal{D}_{drop}$ does not affect their interpretation).

Since $\hat{\beta}^*$ is not a likelihood estimator, it is not strictly valid for $\hat{\beta}^*$ to be used to evaluate likelihood-based diagnostic statistics. However, it is easy to show that it is approximately valid to do so in large samples. Standard likelihood-based test statistics (e.g., Likelihood Ratio Test and Deviance Test) involve evaluating the model log-likelihood $l(\hat{\beta}|\mathcal{D})$, where $\hat{\beta}$ is the standard ML estimator and $\mathcal{D}$ are the microdata used to fit the model. Using a second order Taylor Series approximation to $l(\beta|\mathcal{D})$ centred around $\hat{\beta}$ and noting $\hat{\beta}^* = \hat{\beta} + (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{E}^*$, it follows that $l(\hat{\beta}^*|\mathcal{D}) \approx l(\hat{\beta}|\mathcal{D}) - 3^{-1}trace\{(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\}$ which means for large n that $l(\hat{\beta}^*|\mathcal{D}) \approx l(\hat{\beta}|\mathcal{D})$. Furthermore, if the number of dropped records is small then $l(\hat{\beta}^*|\mathcal{D}_{Drop}) \approx l(\hat{\beta}|\mathcal{D})$. For large n, this means that a standard likelihood-based test statistic evaluated at $\hat{\beta}^*$ and $\mathcal{D}_{drop}$ is approximately the same as a standard likelihood test statistic (i.e., $t^* \approx t$). This approximation is verified in empirical evaluations.

In small samples, it may be worthwhile to adjust some statistics to make them valid. For example, the standard Wald Test statistic is $t_W = \hat{\beta}'(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\beta$ and is distributed as chi-squared with K degrees of freedom. The adjusted Wald statistic is

$$t_W^*(\hat{\beta}^*, \mathcal{D}_{drop}) = \hat{\beta}^{*\prime}\left[(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})_{drop}^{-1} + (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})_{drop}^{-1} Var(\mathbf{E}^*)(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})_{drop}^{-1}\right]\hat{\beta}^*,$$

and is chi-squared with K degrees of freedom.

The only protection of t^* from attacks is that it is calculated from $\mathcal{D}_{drop}$ rather than $\mathcal{D}$. To be consistent with the protections given to regression parameters (see Subsection 3.2), consider the perturbed statistic

$$t^{**} = t^* + e(t)u^*, \tag{7}$$

where $e(t)$ bounds the maximum influence that a record on the microdata has on the statistic t, and u^* is a random variable sampled from the uniform distribution on the range $(-1,1)$. If the same model is fitted then the same value for u^* must be generated (cf. averaging over $\mathbf{E}^*$ s). All the attacks discussed previously on regression parameters can be

reformulated to be attacks on diagnostic statistics, t^{**}. For reasons of space these are not mentioned.

For example, in the case of the dispersion parameter for a logistic regression, ideally $\hat{\phi}^* = (n-K)^{-1}\Sigma_i(y_i - \mu_i^*)^2 v^{-1}(\hat{\mu}_i^*)$ would be released. Since $e(\phi) \approx n^{-1}$, the released dispersion parameter $\phi^{**} = \phi^* + e(\phi)u^* = \phi^* + O(n^{-1})$, which means that perturbation will only have a small impact. Moreover, it is easy to show, using first-order Taylor Series approximations, that for many test statistics $t^{**} = t + O(n^{-1})$ – implying that the difference between the standard and released statistics will be small. This is verified in a limited empirical study.

For statistics used in hypothesis testing, only the ranged p-value for the test statistic, t^{**}, should be reported, rather than the value of the test statistic and the degrees of freedom. The degrees of freedom for t and t^{**} for the above mentioned test statistics are the same, using as justification the fact that $\phi^{**} \approx \phi$. Sparks et al. (2008) suggest reporting the p-values in the ranges $[0, 0.001)$, $[0.001, 0.01)$, $[0.01, 0.05)$, $[0.05, 0.1)$ and $[0.1, 1)$.

The challenge of confidentialising graphical output, including exploratory data analysis, in remote analysis systems is discussed by Sparks et al. (2008) and by many other authors (for a review see O'Keefe and Chipperfield 2013). This however, has not considered the risks from linked data. This is an interesting and useful avenue for future work.

4. Evaluation of Risk and Utility of a Remote Server: Linking the Australian Census to the Migrants Database

The ABS Census of Population and Housing provides economic and social information about migrants living in Australia. However, there are certain questions of great interest about migrants that the Census data alone cannot answer. One key question is how migrant visa class, assigned prior to arrival in Australia, is related to post-arrival social and economic outcomes. The different visa classes include *family, humanitarian, skilled, onshore* and *other*. Answering such a question is made possible through linking the Census with the Department of Immigration and Citizenships (DIAC) Settlement Database (SDB) which collects *visa class*. These answers would assist with the future development and evaluation of immigration programs and support services for migrants.

The Census 2006 microdata are made up of more than 20 million records. The reference period for the Census is 8 August 2006. For this study, the SDB had a reference period from 1 January 2000 to 8 August 2006 (Census night) and contained the records of 861,000 persons who, during that period, were granted visas to live permanently in Australia. DIAC provided the SDB to the ABS for the purpose of linking it with the Census. The variables used to probabilistically link records on the SDB and Census were *age* (in years), *month and day of birth, marital status* (five categories), *sex, country of birth, year of arrival, religion, main language* and *small area geography*. About 530,000 records were linked. For the purposes of this study, the linked file includes select Census variables, the SDB variable *visa class* and the linking variables *age, marital status, sex, country of birth, year of arrival, main language* and *small area geography*. For the purpose of this study we assumed that the linking variables *religion* and *month and day of birth* were not included on the linked data. This means DIAC would have access to seven variables and small area geographic information on the linked microdata. If more SDB

variables were included on the linked microdata, the disclosure risk would likely be greater than that measured below.

In this study the ABS is the IA and Data Custodian T and DIAC is Data Custodian A. The ABS, as an IA, is planning to release the SDB Census-linked microdata through its remote server. The ABS is legally obliged to ensure that the risk of disclosing information about a particular person is *unlikely*. This legislation (Census and Statistics Act 1905) does not distinguish between sensitive and nonsensitive variables and does not make a special provision for *trusted analysts*. (The case study here is an example of a general strategy of the ABS to link its Population Census to microdata collected by select government departments. Details on the legal framework behind an IA in Australia can be found on the ABS website).

Subsection 4.1 considers the utility of modelling with and without the protections of Section 3, and Subsection 4.2 considers the disclosure risk in a high-risk scenario.

4.1. *Empirical Evaluation of Utility*

While there are many possible research questions, one of particular importance to policy makers is to what extent migrants have difficulty finding employment after they arrive in Australia and how this is related to visa class. A useful way to answer such a question is to fit a regression model to employment with a range of covariates, including visa class. Tables 1 and 2 give the results of fitting such a model to two populations- the first is all migrants living in the Australian Capital Territory (ACT) and the second is all migrants living in the ACT who arrived after 2001, respectively. The set of restrictions of Subsection 3.2 did not prevent the models being fitted.

The results show that $\hat{\beta}^*$ with $\phi = 1$ (remembering that ϕ controls the magnitude of the perturbation) and the standard estimator $\hat{\beta}$ were very similar. As mentioned above, the standard errors of $\hat{\beta}^*$ can be computed by using either an analytic or Jackknife expression for the first term in (6). Tables 1 and 2 show that the difference between the two variance estimates is generally small and tends to be larger for coefficients of covariates that have a low frequency. Consequently, the tests for the statistical significance of the regression coefficients were almost identical whether they were based on $\hat{\beta}^*$ with Jackknife standard errors or $\hat{\beta}$ with analytic standard errors. The one exception was in Table 1, where the coefficient $55 < age < 64$ was not statistically significant at the 95% level after the protections were applied. Coefficients of covariates with a low frequency tend to be more influenced by perturbation of the score function. Tables 3 and 4 illustrate that the standard and released diagnostics statistics are very similar. Overall, this section illustrates that the protections had only a small impact on inference.

4.2. *Simulated Evaluation of Risks*

This section simulates attacks that could be conducted by an analyst with access to the SDB. The aim of such simulated attacks is to infer the value of one or more Census variables, using statistical output released by the remote server and the SDB. While the simulation does not involve use of the DIAC Census-linked microdata, it aims to replicate the possible attacks on the linked microdata. The benefit of simulation is that it is

Table 1. Impact of Protections on Regression Coefficients (ACT)

Variable name	Frequency ($n=$)	$\hat{\beta}$	$\hat{\beta}^*$	Analytic Standard Error of $\hat{\beta}$	Analytic Standard Error of $\hat{\beta}^*$	Jackknife Standard Error of $\hat{\beta}^*$
constant	5,161	−1.13	−1.06	0.19	0.20	0.19
school qual.	303	0.39	0.41	0.15	0.15	0.13
female	2,825	1.14	1.14	0.08	0.08	0.08
tertiary qual.	4,045	−0.56	−0.55	0.09	0.09	0.09
part-time student	493	−0.20	−0.20	0.14	0.14	0.14
full-time student	935	2.00	1.99	0.10	0.10	0.10
non-urban	25	−0.11	−0.11	0.60	0.61	0.79
not married	1,798	−0.39	−0.42	0.10	0.10	0.08
family visa	2,197	0.40	0.39	0.08	0.08	0.08
humanitarian visa	191	0.56	0.57	0.19	0.19	0.17
other visa	49	0.35	0.48	0.39	0.38	0.41
onshore visa	2,162	−0.19	−0.19	0.08	0.08	0.08
English spoken at home	1,324	−1.18	−1.21	0.15	0.15	0.17
English proficient	3,458	−0.75	−0.77	0.13	0.13	0.16
$25 \leq age \leq 34$	2,331	−0.08	−0.13	0.12	0.12	0.12
$35 \leq age \leq 44$	1,456	−0.05	−0.11	0.14	0.14	0.14
$45 \leq age \leq 54$	506	−0.24	−0.30	0.18	0.18	0.17
$55 \leq age \leq 64$	116	0.54	0.46	0.26	0.26	0.27

Table 2. Impact of Protections on Regression Coefficients (ACT and Year of Arrival Prior to 2001)

Variable name	Frequency ($n=$)	$\hat{\beta}$	$\hat{\beta}^*$	Analytic Standard Error of $\hat{\beta}$	Analytic Standard Error of $\hat{\beta}^*$	Jackknife Standard Error of $\hat{\beta}^*$
constant	1,529	−1.51	−1.42	0.46	0.50	0.56
school qual.	88	0.83	0.86	0.27	0.28	0.27
female	825	1.30	1.31	0.18	0.18	0.19
tertiary qual.	1,226	−0.75	−0.74	0.19	0.19	0.16
part-time student	156	−0.29	−0.31	0.29	0.29	0.28
full-time student	134	1.97	1.93	0.26	0.26	0.19
non-urban	11	0.92	1.49	1.14	1.37	0.99
not married	481	−0.45	−0.45	0.20	0.20	0.22
family visa	727	0.58	0.57	0.18	0.18	0.18
humanitarian visa	44	1.06	1.10	0.40	0.40	0.44
other visa	38	0.01	−0.17	0.53	0.55	0.80
onshore visa	795	0.10	0.10	0.16	0.16	0.17
English spoken at home	422	−1.05	−1.07	0.32	0.32	0.32
English proficient	1,032	−0.95	−0.95	0.30	0.30	0.29
$25 \leq age \leq 34$	583	−0.30	−0.39	0.33	0.33	0.36
$35 \leq age \leq 44$	587	0.09	0.00	0.34	0.34	0.33
$45 \leq age \leq 54$	181	−0.05	−0.14	0.39	0.42	0.38
$55 \leq age \leq 64$	27	0.68	0.39	0.59	0.61	0.66

Table 3. Impact of Statistical Disclosure Control on Diagnostic Statistics (ACT)

Statistic	Standard (t)	95% interval for t^{**}
Dispersion, ϕ	0.93	(0.92, 0.93)
$R-square$	0.18	(0.18, 0.18)
Likelihood Ratio	1052 (<0.001)	(1035, 1057) (<0.001)$^{\psi}$

$^{\psi}$only the ranged p-value is released.

possible to readily construct a situation that both is realistic and presents a high risk of disclosure.

The ABS, as an IA, would not reveal to data custodians which records were linked (e.g., in the Census-SDB linkage only 530,000 of the 861,000 SDB records were linked). However, it is assumed in this simulation that the attacker could identify a specific subpopulation of records that are very likely to be linked correctly. For example, in the Census-SDB linkage it may be inferred that certain subpopulations of records (e.g., proficient in English and high level of education) have a very high chance of reliably reporting linking variables, and so are likely to be linked correctly to their corresponding Census records.

4.2.1. Simulated Subpopulation

Assume the attacker fits models to a subpopulation of the linked microdata of size $n = 30$ or 50 records. This subpopulation could be defined in terms of small area geography, available on the SDB. Given the previous assumption, the attacker knows the exact set of records in the subpopulation. To make this simulation realistic, the attacker chooses to use eight variables on the linked microdata: small area geography to define the subpopulation of size n, the six other remaining SDB variables (see above), denoted by $\mathbf{x}$, and one Census variable (e.g., employment), denoted by y. In the notation of Section 2, the attacker knows $\mathbf{X}$ and seeks to infer y_i for some or all i. The variables for records in the subpopulation were independently generated 200 times in the following way:

- Each record has a unique covariate pattern in $\mathbf{x}$. Since $\mathbf{x}$ has dimension six, there are $2^6 = 64$ possible covariate patterns, of which $n = 30$ or 50 are randomly selected for the subpopulation.
- $S_y = \Sigma_i y_i = 3,6$ where y is generated from the logistic model $1/(exp(-\eta_i))$, $\eta_i = 1.6 + x_{1i} - 1.5x_{2i} + 1.3x_{3i} - 0.8x_{4i} + 1.3x_{5i} + 0.9x_{6i} + e_i$ and the e_i s are independent standard normal random variables. These model parameters were chosen arbitrarily but to be within the range of those in Tables 1 and 2 and to generate the desired range in S_y.

Table 4. Impact of Statistical Disclosure Control on Diagnostic Statistics (ACT and Year of Arrival Prior to 2001)

Statistic	Standard (t)	95% interval for t^{**}
Dispersion, ϕ	0.93	(0.91, 0.94)
$R-square$	0.18	(0.18, 0.19)
Likelihood Ratio	257 (<0.001)	(240, 261) (<0.001)$^{\psi}$

$^{\psi}$only the ranged p-value is released.

Since each value for x is unique and SDB contains the name and address for every record, disclosure automatically occurs if the attacker who has access to the SDB is able to infer the value of y_i for any $\mathbf{x}_i$. This is because there is a 1-1 correspondence between $\mathbf{x}_i$ and name and address for all i.

The ABS releases frequency counts from its Census microdata via its remote server. While a small amount of noise is added to these counts before they are released, it is frequently assumed in this simulation that S_y is in the public domain. This is a strong assumption since, as mentioned above, such counts are perturbed by a small amount.

In reality, it is unlikely all of the above conservative assumptions made for this simulation will be true. As a result, the disclosure risks would in reality be significantly lower than those measured in this section.

4.2.2. Attacks Using Regression Coefficients

The effectiveness of two attacks were measured on the 200 independently simulated subpopulations. It was interesting to see how the success of an attack was influenced by whether the remote server released:

- $\hat{\beta}$. This effectively means there is no (N) protection.
- $\hat{\beta}^*(\mathcal{D})$ computed from (3) but using $\mathcal{D}$ instead of D_{drop}. The protection is from perturbation (P) of the score function.
- $\hat{\beta}^*(\mathcal{D}_{drop})$ computed from (3). The protection is from perturbation and dropping a single randomly selected record (O,P), where O denotes dropping.

The first attack was *Solving the Estimating Equation* (SEE) (see *Example 1* and Subsection 3.3). When the remote server uses the O and P protections, SEE involved finding all possible values for $\mathbf{y}$ that are solutions to (4) given S_y, $\mathbf{X}$ and $\hat{\beta}^{*(m)}$ for $m = 1, \ldots, M$. Disclosure occurred for record j if, across all possible solutions, the value for y_j was always unique. Table 5 gives the proportion of SEE attacks that were successful in a range of scenarios. For example, Table 5 shows that when $n = 50$ and there were no protections, all values in $\mathbf{y}$ were disclosed in every one of the 200 simulated subpopulations from only a single model; if instead the P protection was used with $\phi = 1$, the success rate fell to 2%. A summary of the findings from Table 5 are described below.

- Releasing $\hat{\beta}$ was a high disclosure risk. The risk was 100% when y was the dependent variable.
- As ϕ increased the success rate reduced. However, the P protection on its own did not reduce the success rate to zero.
- The success rate increased as M, the number of fitted models, increased.
- The O protection on its own did not reduce the success rate.
- If only the P protection was used, uncertainty in S_y (see 6^{ψ} in Table 5) did not seem to provide much protection.
- If both the P and O protections were used, the disclosure risk was zero.

The second attack was Differencing Counts (DC) (see Subsection 2.1.3 and Subsection 3.3). The target record for a differencing attack was chosen completely at

Table 5. The Success Rate of Solving the Estimating Equation (SEE)

Number of Models	Dependent variable(s)	Defence[ψ/ψ]	n	S_y	Percentage of Attacks which inferred $y = 0$ for at least one record	Percentage of Attacks which inferred $y = 1$ for at least one record	Percentage of Attacks which inferred y for all records
1	y	N	30	6	100	100	100
1	x_1	N	30	6	93	42	10
1	y	N	50	6	100	100	100
1	x_1	N	50	6	82	9	3
1	y	O	30	6	100	100	100
1	y	O	50	6	100	100	100
1	y	P ($\phi = 1$)	30	6	16	0	0
1	x_1	P ($\phi = 1$)	30	3	92	12	2
1	x_1	P ($\phi = 1$)	30	6	76	12	0
1	x_1	P ($\phi = 1$)	50	3	86	4	0
1	x_1	P ($\phi = 1$)	50	6	44	0	0
3	y, x_1, x_2	P($\phi = 1$)	30	6	93	40	5
5	y, x_1, x_2, x_3, x_4	P ($\phi = 1$)	30	6	90	73	9
7	all	P ($\phi = 1$)	30	6	97	49	12
7[ψ]	all	P ($\phi = 1$)	30	6[ψ]	82	33	7
1	y	P ($\phi = 2$)	30	6	2	0	0
1	x_1	P ($\phi = 2$)	30	6	0	0	0
1	x_1	P ($\phi = 2$)	30	3	0	0	0
1	x_1	P ($\phi = 2$)	50	3	0	0	0
1	x_1	P ($\phi = 2$)	50	6	0	0	0
3	y, x_1, x_2	P($\phi = 2$)	30	6	0	0	0
5	y, x_1, x_2, x_3, x_4	P ($\phi = 2$)	30	6	43	8	2
7	all	P ($\phi = 2$)	30	6	69	78	2
3	y, x_1, x_2	O,P ($\phi = 1$)	30	3	0	0	0
5	y, x_1, x_2, x_3, x_4	O,P ($\phi = 1$)	30	3	0	0	0
7	all	O,P ($\phi = 1$)	30	3	0	0	0

[ψ] while $S_y = 6$ the attacker only knew $S_y = 5$, 6 or 7.

[ψ/ψ] N – No protection, O – Dropping one record completely at random, P – Perturbing.

Table 6. Differencing Attack

Defence	S_y	Success Rate (%)
N	30	100
P	30	5
O,P	30	0

random. Table 6 shows that the proportion of differencing attacks that were successful when protections N, P and (P and O) were used was 100%, 5% and 0% respectively.

For the results in Table 5, $L_{DIAC} = 0.5$ (see Subsection 3.1 where A = DIAC) for fitting a single model and $L_{DIAC} = 3.5$ when seven models were fitted. By contrast, for the models in Tables 1 and 2, $L_{DIAC} = 0.001$ and 0.002 respectively; these values are considerably smaller since most variables in the model were not SDB variables and the sample size was larger. Interesting further work would identify the optimal value for L_{DIAC} to trigger an audit by the ABS. If $L_{DIAC} > 1$ was to trigger such an audit, the audit would readily identify that the fitted models have the distinctive feature of the SEE attack (see Subsection 2.1.2). Remedial action could then be taken by DIAC and ABS to prevent further attacks.

The ABS, as an IA, could consider dropping variables from the linked microdata that are common to Census and SDB. If a common variable has limited analytic value, the ABS, as the IA, should consider dropping it from the linked microdata. This is particularly the case if a common variable is useful in uniquely identifying a record. Dropping such variables will limit the prior knowledge, and hence the effectiveness, of an attack.

5. Discussion

Modern advances have allowed vast amounts of microdata to be collected by data custodians. With increasing sophistication of policy makers and the consequent demand for more detail, linking such microdata across data custodians is becoming increasingly important. While the benefits to society of allowing access to linked microdata are significant, data custodians need to ensure that allowing access is unlikely to result in the disclosure of information about a particular person or organisation. The Australian Bureau of Statistics (ABS) is playing a lead role in developing a framework for the integration of Australian Commonwealth data. The role of an Integrating Authority (IA) is to maximise the inherent value of Commonwealth data to society, to facilitate access to the linked data and to ensure disclosure risk is acceptable. The ABS is developing infrastructure in the areas of record linkage and remote analysis to support its goal to become the lead IA in Australia.

This article proposes a set of protections that an IA can apply to statistical output from linked microdata. The evaluations show that the protections prevent disclosure in a high-risk scenario and have only a small impact on inferences for analysis involving moderate sample sizes. The method in the article can be readily extended to three or more data custodians. Importantly, this article shows that some popular protections against disclosure (e.g., dropping records, rounding regression coefficients or imposing restrictions on model selection) are perhaps not as effective as previously thought.

There is a need to extend the approach here to include analysis of continuous variables. Extensions to multilevel models is also important, since linked administrative data are often longitudinal in nature or contain a natural hierarchy.

6. References

Bleninger, P., Drechsler, J., and Ronning, G. (2010). Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study. Privacy in statistical databases, J. Domingo-Ferrer and E. Magkos (eds). New York: Springer.

Chambers, R.L. and Skinner, C.J. (2003). Analysis of Survey Data. Hoboken, NJ: John Wiley and Sons.

Churches, T. and Christen, P. (2004). Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making, 4, Available at: http://www.pubmedcentral.nih.gov/tocrender.fcgi?iid=10563 (accessed June 2012).

Cox, L.H., Karr, A.F., and Kinney, S.K. (2011). Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think but not How to Act. International Statistical Review, 79, 160–183. DOI: http://dx.doi.org/10.1111/j.1751-5823.2011.00140.x

Dwork, C. and Smith, A. (2009). Differential Privacy for Statistics: What We Know and What We Want to Learn. Journal of Privacy and Confidentiality, 1, 135–154.

Gomatam, S., Karr, A., Reiter, J., and Sanil, A. (2005). Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Systems. Statistical Science, 20, 163–177. DOI: http://dx.doi.org/10.1214/088342305000000043

Herzog, T.N., Scheuren, F.L., and Winkler, W.E. (2007). Data Quality and Record Linkage. Berlin: Springer.

Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression. Hoboken, NJ: John Wiley and Sons Inc.

Karr, A.F., Lin, X., Sanil, A.P., and Reiter, J.P. (2009). Privacy Preserving Analysis of Vertically Partitioned Data Using Secure Matrix Products. Journal of Official Statistics, 25, 125–138.

Kohnen, C. and Reiter, J.P. (2009). Multiple Imputation for Combining Confidential Data Owned by Two Agencies. Journal of the Royal Statistical Society Series A, 172, 511–528. DOI: http://dx.doi.org/10.1111/j.1467-985x.2008.00574.x

Lucero, J. and Zayatz, L. (2010). The Microdata Analysis System at the U.S. Census Bureau. Privacy in Statistical Databases, J. Domingo-Ferrer and E. Magkos (eds). New York: Springer.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models (2nd ed.). London: Chapman and Hall.

O'Keefe, C. and Chipperfield, J.O. (2013). A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems. International Statistical Review. DOI: http://dx.doi.org/10.1111/insr.12021

O'Keefe, C. and Good, N. (2009). Regression Output from a Remote Analysis System. Data & Knowledge Engineering, 68, 1175–1186. DOI: http://dx.doi.org/10.1016/j.datak.2009.06.009

O'Keefe, C., Sparks, R., McAullay, D., and Loong, B. (2012). Confidentialising the Output of a Survival Analysis in a Remote Analysis System (to appear). Journal of Privacy and Confidentiality, 4, 127–154.

Reiter, J. (2002). Satisfying Disclosure Restrictions with Synthetic Data Sets. Journal of Official Statistics, 18, 511–530.

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. Hoboken, NJ: John Wiley and Sons.

Shlomo, N. (2007). Statistical Disclosure Control Methods for Census Frequency Tables. International Statistical Review, 75, 199–217. DOI: http://dx.doi.org/10.1111/j.1751-5823.2007.00010.x

Skinner, C. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-Linear Models. Journal of American Statistical Association, 103, 989–1001. DOI: http://dx.doi.org/10.1198/016214507000001328

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C., Duncan, J., and Keighley, T. (2008). Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics™. Computer Methods and Programs in Biomedicine, 91, 208–222.

Received January 2013
Revised October 2013
Accepted November 2013