Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers

S. Gomatam[*], A. F. Karr[†], J. P. Reiter[‡], and A. P. Sanil[§]

November 29, 2004

Abstract

Given the public's ever-increasing concerns about data confidentiality, in the near future statistical agencies may be unable or unwilling, or even may not be legally allowed, to release any genuine microdata—data on individual units, such as individuals or establishments. In such a world, an alternative dissemination strategy is remote access analysis servers, to which users submit requests for output from statistical models fit using the data, but are not allowed access to the data themselves. Analysis servers, however, are not free from the risk of disclosure, especially in the face of multiple, interacting queries. We describe these risks and propose quantifiable measures of risk and data utility that can be used to specify which queries can be answered, and with what output. The risk-utility framework is illustrated for regression models.

Key words: Data confidentiality, data utility, disclosure risk, microdata, regression server, remote access server, statistical disclosure limitation

1 Introduction

When disseminating microdata—on individual units, such as people or establishments—to the public, to researchers or to other agencies, national statistical agencies face conflicting missions. They seek to release microdata that support a wide range of statistical analyses, yet they also must safeguard the confidentiality of respondents' identities and attribute values. Agencies that fail to protect confidentiality may face serious consequences. They, or their employees, may be be subject to legal actions. They may lose the trust of the public, so that respondents are less willing to participate in studies or to provide accurate data.

Even when identifiers such as names and addresses or social security numbers are removed before releasing data, there remain serious risks of disclosure. For example, ill-intentioned users ("intruders") may be able to link released records to external databases, which are proliferating at all levels of government as well as in the private sector. For example, many towns and cities sell or make available on-line databases containing voter registrations, and Sweeney (1997) showed that 97% of the records in a medical database for

[*]National Institute of Statistical Sciences, Research Triangle Park, NC, USA. Now at the US Food and Drug Administration.
[†]National Institute of Statistical Sciences, Research Triangle Park, NC, USA.
[‡]Duke University, Durham, NC, USA
[§]National Institute of Statistical Sciences, Research Triangle Park, NC, USA.

Cambridge, MA could be identified by birth date and 9-digit ZIP code by linking them to a voter registration list.

To reduce disclosure risks, agencies typically alter the original data before release, for example by perturbing, coarsening or swapping data values (Willenborg and de Waal, 2001). Of course, such statistical disclosure limitation (SDL) techniques also reduce the usefulness of the released data.

As more external databases become available and record linkage technologies improve, it becomes virtually mandatory to contemplate a world in which useful microdata releases are no longer feasible. In a world without microdata, three approaches to dissemination remain viable. The first and simplest is to release only data summaries such as low-dimensional tables, graphs and maps. Such summaries are less useful in some contexts than complex analyses, and there remain disclosure risks. For example, cell counts in a table can be bounded, possibly very accurately, from released marginal totals (Dobra et al., 2002, 2003).

The second approach is to release synthetic—that is, simulated—microdata (Rubin, 1993). Synthetic databases can have low disclosure risks, since some or all of the released values are not genuine, but this also decreases utility of the data. Both risk and utility depend strongly on the model used for synthesis. See Little (1993), Fienberg et al. (1996, 1998), Raghunathan et al. (2003), and Reiter (2002, 2003a, 2004) for further discussion.

The third approach, which is the subject of this paper, is to release the results of statistical analyses of the data, such as estimated model parameters and standard errors, without releasing any microdata. This approach can be implemented using remote access analysis servers, to which users submit requests for analyses and, in return, receive some form of output (Keller-McNulty and Unger, 1998; Duncan and Mukherjee, 2000; Schouten and Cigrang, 2003). In a world without microdata, the analysis dissemination approach has advantages over the other two approaches. It permits a wider range of analyses than does releasing only data summaries, and it provides results based on actual rather than simulated microdata. Several statistical agencies are developing or already use servers as part of their data dissemination strategies, including the Australian Bureau of Statistics, Statistics Canada, Statistics Denmark, Statistics Netherlands, Statistics Sweden, the US Census Bureau, the US National Agricultural Statistics Service, the US National Center for Education Statistics, and the US National Center for Health Statistics (Rowland, 2003).

Even though they prevent direct access to the data, analysis servers do not preclude disclosures. It may be possible for intruders to learn identities or attribute values by means of "targeted" queries. Furthermore, queries that are innocuous individually may produce disclosures collectively. Because of these possibilities, we believe it is necessary to formulate a risk-utility framework (Duncan et al., 2002), based on quantified measures of disclosure risk and data utility, for deciding in a principled way which queries can be answered by analysis servers. In this paper, we present such a framework, with an initial, specific application to servers that disseminate the results of linear regression analyses.

The remainder of the paper is organized as follows. §2 contains background on disclosure risk and SDL techniques. §3 describes the statistical components of analysis servers. §4 suggests how users successfully can perpetrate disclosure attacks on servers, as well as methods for limiting the success of these attacks. §5 presents quantitative measures of risk and utility for servers, illustrating their use with simulations of regression modeling. §6 concludes with an agenda for future research.

2 Background on SDL

This section is a primer on statistical disclosure limitation. See Duncan and Lambert (1986), Federal Committee on Statistical Methodology (1994), Paass (1988), Willenborg and de Waal (1996) and Willenborg and de Waal (2001) for further information.

There are three principal forms of disclosure for microdata (Lambert, 1993). Identity disclosure occurs when a record in the database can be associated with the individual unit it describes. Attribute disclosure occurs when the value of a sensitive attribute, such as income or health status, is disclosed directly.

Inferential disclosure, the principal risk addressed in this paper, occurs when units are threatened not by their records but by statistical characteristics of the entire database. For example, suppose that automobile operating expenditures, which seem innocuous, are a good predictor of medical expenditures, which are not innocuous. In some locales, such as rural areas that entail significant travel to reach medical centers and where there is no public transportation, this is at least plausible. If this relationship were known *and known to be a good relationship*, an intruder with access to travel expenditures could predict medical expenditures. Another example (Palley and Simonoff, 1987) occurs for business data. Organizations may want relationships between salaries and non-confidential variables to be protected, because otherwise, some employee could fit a model that reveals his or her salary is less than predicted.

For inferential disclosure, the mere existence of some relationship may threaten confidentiality, but more often the threat is in the quantitative details and the strength of the relationship. For example, it is obvious that household income, a natural attribute to protect, is positively correlated with home value, which in most jurisdictions is public information. No one can be prevented from "knowing" that the relationship exists, but the values of either regression coefficients (the quantitative details) or the correlation (the strength of the relationship) may be suppressed in the name of SDL.

To protect data confidentiality and meet users' demands for microdata, agencies and researchers have developed an array of SDL strategies (Duncan et al., 1993). At the highest level, SDL divides into strategies based on restricted access and those based on restricted data. Mechanisms for restricted access include data centers, licensing, and vetting of researchers and their research plans. Restricted access SDL strategies allow users to perform analyses directly on the underlying data, although specific analyses may be suppressed, either *a priori*, if the analysis is known to threaten confidentiality, or *a posteriori*, the output reveals a threat. These centers rely on the honesty of researchers to protect confidentiality, and can be expensive for agencies and inconvenient for researchers.

Restricted data SDL strategies alter the data in ways that limit potential for disclosure. For example, the first step in preventing identity disclosures is to remove explicit identifiers such as name, address and social security number, as well as implicit identifiers, such as "Occupation = Mayor of New York." Almost always, however, this is not enough. Again a broad bifurcation occurs: restricted data strategies either produce information releases, such as tabular summaries and statistical analyses of the data, or data-like releases. Analysis servers are an example of a restricted data, information release SDL strategy. Restricted data, data-like SDL strategies include aggregating or coarsening the underlying microdata. For example, to protect units with high incomes, income is frequently "top-coded," so that one category is "More than $X." They also include perturbing original values, such as by swapping data (Dalenius and Reiss, 1982; Gomatam et al., 2003) or adding random noise to units' values (Fuller, 1993).

Restricted data SDL strategies can be applied with varying intensity. The amount of information released can be limited to subsets of varying sizes; aggregation may be relatively fine or very coarse; relatively few or rather many data values may be swapped or perturbed. Generally, the higher the SDL intensity, the greater the protection against disclosure risk, but the less the utility of the released data.

At least implicitly, agencies choose SDL strategies by balancing confidentiality protection and utility of the released information. We advocate use of explicit risk-utility frameworks to choose SDL strategies, as proposed by Duncan et al. (2002). The general idea is to quantify the disclosure risk and data utility of possible SDL strategies, and then select strategies that give the highest utility for acceptable confidentiality protection. Explicit approaches have been applied successfully in a variety of settings (Dobra et al., 2002,

2003; Domingo-Ferrer et al., May, 2001; Gomatam et al., 2003; Yancey et al., 2002).

Entirely different sets of issues and strategies arise when analyses involve distributed databases that cannot actually be integrated (Karr et al., 2004b,a; Sanil et al., 2004b,a).

3 Description of Analysis Servers

Explicit risk and utility measures have not been developed for analysis servers. To begin our development of such measures, we define the statistical components of analysis servers.

3.1 Conceptual Framework

Let $\mathcal{D}$ be the microdata collected by the agency, either through a survey or census. A *server* is a software system that releases functions of the data, $F(\mathcal{D})$. These functions might include visualizations, estimates and summaries of distributions of variables, or estimates of functional relationships among variables using complex statistical models. The server receives from the user a query Q for some $F(\mathcal{D})$, and it responds either by providing $F(\mathcal{D})$ or refusing to do so because of confidentiality or utility considerations. A more complex response strategy would be to provide an alternative analysis rather than a refusal.

In addition to $\mathcal{D}$, the components of the server include:

- *Query space*, the set $\mathcal{Q}$ of queries that the server can process. For example, some servers can handle requests for tabular data analyses but not regression analyses, whereas others do the opposite. The server responds to any $Q \in \mathcal{Q}$ with either the requested $F(\mathcal{D})$ or a refusal to provide $F(\mathcal{D})$.

- *Answer space*: This is the set $\mathcal{A} \subseteq \mathcal{Q}$ of queries that the server answers with statistical output. We assume that the query for $F(\mathcal{D}) = \mathcal{D}$ is never answered.

- *Disclosure risk measure*, a real-valued function such that $R(Q_1, \ldots, Q_m)$ is the disclosure risk of providing $F(\mathcal{D})$ for the set of queries $\{Q_1, \ldots, Q_m\}$.

- *Data utility measure*, a real-valued function such that $U(Q_1, \ldots, Q_m)$ is the data utility of providing $F(\mathcal{D})$ for the set of queries $\{Q_1, \ldots, Q_m\}$.

The risk and utility measures are the components of a *query mediation mechanism* that determines $\mathcal{A}$. The query mediation mechanism must address the problem of interaction among queries: answering several queries may allow users to piece together enough information to achieve disclosures. This issue has been recognized by several authors (Palley and Simonoff, 1987; Duncan and Mukherjee, 2000; Dobra et al., 2002), and is discussed further in §4.

Servers may be either static or dynamic. In a static server, $\mathcal{A}$ is pre-computed. The underlying query mediation mechanism is typically based on either (i) optimization of $U(\mathcal{A})$ subject to an upper bound constraint on $R(\mathcal{A})$; or (ii) selection of $\mathcal{A}$ from a frontier of undominated candidate spaces $\mathcal{A}^c$, i.e., those for which no other candidate release has both lower disclosure risk and higher data utility. Both of these query mediation mechanisms are illustrated in §5.

Dynamic servers accept queries in real time and respond expeditiously if not immediately; $\mathcal{A}$ is determined by the queries that the server elects to answer. Ultimately, a dynamic server reaches a terminal state in which no remaining unanswered queries are answerable. The disclosure risk and data utility associated with responding to a query must take into account those queries that have been answered previously. Dynamic

servers present challenges at multiple levels. Practical issues include scalable computational implementations. Conceptual issues include abstractions such as accounting for the fact that each answered query makes others unanswerable. There are policy issues as well, notably user equity, to prevent a single user or group of users from exerting undue influence on the trajectory of the system. Whether dynamic servers are possible remains an open question.

One relatively well understood class of servers is *table servers* (Dobra et al., 2003, 2002; Karr et al., 2003). In this case, $\mathcal{D}$ is a large contingency table containing counts or sums, $\mathcal{Q}$ is a partially ordered set of marginal sub-tables of $\mathcal{D}$, and responses are either the requested sub-table or refusal. Even in this relatively simple case, computational and policy issues are challenging.

We assume that the metadata associated with $\mathcal{D}$ are available to users, either directly from the server or through other sources. These metadata include attribute definitions, sample sizes, survey frames, response rates, representations of missing values and similar information.

3.2 Model Servers

In the remainder of this paper, we focus on servers for which the query space $\mathcal{Q}$ consists of requests for relevant output from statistical models involving a response and one or more predictor variables in $\mathcal{D}$. We term these *model servers*. Responses, when not refusals, consist minimally of point estimates of the model coefficients, the estimated covariance matrix of the coefficients, and some global goodness-of-fit measures, such as coefficients of determination R^2, dispersion parameters and deviances. We also assume that the means and standard deviations of all variables in $\mathcal{D}$ are available.

These assumptions are not without import. In particular, we believe strongly that a model should never be released without at least global measures of fit, and in most cases, as we discuss in the next paragraph, local measures of fit. Moreover, in most cases, utility considerations would militate against release of a "bad" model. Therefore, a released model can be, in the hands of an intruder, a significant threat to confidentiality.

Ideally, the response from the server should include some way for users to check the fit of models. Obviously, releasing the usual, unit-specific diagnostic statistics can disclose data values. For example, when actual residuals and predicted values are released for a submitted linear regression model, the user can obtain the values of the response by simply adding the residuals to the predicted values.

For diagnosing some types of assumption violations, however, the exact values of the residuals and independent variables are not needed. Rather, the relationships among the residuals and independent variables are examined for patterns in hopes of identifying model mis-specifications. Thus, for remote servers it may be adequate to mimic patterns in the real-data diagnostics without releasing real-data values (Reiter, 2003b; Reiter and Kohnen, 2004). For linear regression diagnostics, the basic idea is to release values of residuals and independent variables simulated from distributions that approximate the relationships between the real-data residuals and independent variables. Users then can treat these synthetic values like ordinary diagnostics quantities, examining scatter plots of the synthetic residuals versus the synthetic independent variables.

4 Disclosures in Model Servers

Our discussion of disclosures is primarily in the context of linear regression modeling, although much of it applies to other models as well. §4.1 describes potential identity and attribute disclosures, and §4.2 describes potential inferential disclosures.

4.1 Identity and Attribute Disclosures for Linear Regressions

By not releasing microdata, and not releasing real-data diagnostics such as residuals, many threats to attribute and identity disclosure are eliminated. However, other threats remain.

In particular, denial of access to microdata does prevent identity and attribute disclosures effected by transformations of variables. Transformation attacks can be used to attempt attribute disclosures when the outcome is a sensitive variable, and to attempt identity disclosures when outcome is a key identifier. The success of these attacks depends on the user's knowledge that certain units with unique values of predictors are in the database, and knowledge of these values. For some databases, such detailed knowledge will not be available, so that disclosures of individuals from transformations may not be likely. However, given the proliferation of publicly available data, it is prudent to assume such knowledge is in the hands of intruders.

Because few operating model servers exist, and those that do exist to our knowledge do not permit transformations, transformation attacks are primarily hypothetical at this point, but they could be simulated on a prototype server, although this would require modeling of intruder knowledge and behavior (Fienberg et al., 1997).

To illustrate, units with unusual values of predictor variables—leverage points—can have a strong effect on the estimated regression, often resulting in small residuals for these units. An intruder who knows that a certain unit is in the database may be able, through transformations, to create artificially extreme leverage points, and thereby learn the outcome variable for that unit from the predicted value of the fitted regression. As an example, suppose X_0 is a sensitive variable unknown to the intruder who also knows that a certain unit m in the database has an unusual value $X_m = x$. The intruder could fit the regression of X_0 on a simple transformation of X_m to increase unit m's leverage, for example by using $1/(|X_m - x| + \epsilon)$ or $\log(|X_m - x| + \epsilon)$, where ϵ is a small positive constant, or by using e^{X_m} when x is large. Transformation of X_0 (e.g., fitting a regression with e^{X_0} as the outcome variable) can further increase the influence of leverage points.

Units need not be leverage points to be subject to transformation attacks. "Dummy variables" can isolate points with unique predictor values. For instance, an intruder who knows a unique predictor value x exactly can learn the associated response by including the predictor $I(X_m = x)$ (or by fitting two regressions, one with $I(X_m \leq x)$ and the other with $I(X_m \leq x - \delta)$ where δ is a small constant).

For categorical predictors, disclosures can occur when there are insufficient numbers of data cases in the categories. For example, an intruder could fit interactions among several categorical variables, such that some cross-classifications describe only one unit. For those cross-classifications, the outcomes can be learned exactly from the fitted values of the regression.

To mitigate the effects of transformation attacks, agencies can limit the space of transformations and types of models that users can submit as queries, but this also reduces data utility. Their effective limitations on transformations should have minimal impact on analyses of interest while satisfactorily controlling disclosure risk. It is also desirable to specify limitations that can be enforced automatically by the server; performing manual checks of every proposed analysis can be time-consuming and expensive.

Next we propose some simple ways whereby agencies can build limitations into model servers. Not all may be useful in any particular context, but they may help prevent classes of transformation attacks with potentially acceptable reductions in data utility. First, key identifiers, such as age, race, and sex, can be prohibited as outcome variables but permitted as predictors. This strategy eliminates identity disclosure attacks that use keys as outcomes. The reduction in data utility can be small, since typically identifiers are not of interest as outcomes. Second, SDL strategies for tabular data can be applied to categorical data in model servers. For example, agencies can prohibit indicator variables from being predictors unless at

least three units with non-identical outcome values satisfy the conditions described by the indicator. The reduction in utility may be small in many data sets, since usually few strong conclusions can be made for units in very sparsely populated categories. Third, transformations that split continuous variables into categories can be disallowed, thereby eliminating attacks that rely on such splits. For servers that permit generalized additive modeling or other methods of curve fitting, this may not substantially sacrifice data utility. Fourth, for any X_i, transformations of the form $g(X_i - h(X_i))$ can be disallowed for all $h(X_i)$ except $h(X_i) = 0$. This prohibits transformations designed to give individual values high leverages. Many transformations for analytical purposes, such as $g(X_i) = \log(X_i)$ or $g(X_i) = \sqrt{X_i}$ do use $h(X_i) = 0$, and so remain permissible. Agencies might allow certain $h(X_i)$, in particular $h(X_i) = \bar{X}_i$, when they are innocuous. Finally, transformations can be disallowed when they increase the leverage values of units, or the values of the X_i, beyond administrator-defined cutoffs. The cutoffs should be set to permit common transformations while preventing outlandish ones whose main purpose is transformation attacks.

Agencies can inform users about the limitations imposed on the answer space, although it may be wise not to disclose cutoff values. Some limitations, like those in the first four points above, can be enforced by the server before submitted models are even fit. Other limitations, like the fifth one above, may have to be enforced dynamically by the server.

4.2 Inferential Disclosures for Linear Regressions

For some databases, agencies may seek to prevent users from fitting particular regressions. For example, an agency may not want to release the output from regressions that have small root mean squared errors and sensitive dependent variables. Or, an agency may want to protect a certain relationship in the data. In this section, we discuss ways that intruders can learn about unreleased regressions through released regressions, and thereby attempt inferential disclosures.

To fix ideas, we define notation used throughout the remainder of the paper. Let $X = (X_0, X_1, \ldots, X_d)$ be the $(d + 1)$ variables in the database $\mathcal{D}$. For any subset $B = \{i, j, \ldots\}$ of variable indices, let $X_B = \{X_i, X_j, \ldots\}$. We write the linear regression of X_a on the predictors whose indices are in B as $X_a|X_B$. For example, the regression X_0 on (X_1, X_2, X_3) is written as $X_0|X_{\{1,2,3\}}$. We use the notation X_{aB} to denote the collection of variables in $X_a \cup X_B$.

Let $\mathbf{X}$ denote the $n \times (d + 1)$ matrix constituting the data for the variables X (n is the number of data cases). For simplicity, we assume that $\mathbf{X}$ has been centered: we use $\mathbf{X}_i - \bar{\mathbf{X}}_i$ for each variable i. Then for any $X_a|X_B$, the vector of least squares estimates of coefficients is

$$\mathbf{b}_{a|B} = (\mathbf{X}_B^t \mathbf{X}_B)^{-1} \mathbf{X}_B^t \mathbf{X}_a.$$

Any $\mathbf{b}_{a|B}$, as well as its estimated covariance matrix and the coefficient of determination $R^2_{a|B}$, can be computed from the sample cross-product matrix

$$\mathbf{S}_{aB} = (\mathbf{X}_a, \mathbf{X}_B)^t (\mathbf{X}_a, \mathbf{X}_B).$$

Hence, a user who obtains $\mathbf{S}_{aB}$ completely from a set of released regressions learns all possible linear regressions involving X_{aB}.

Suppose the server seeks to prevent intruders from learning the coefficients of some sensitive regression, say $X_a|X_B$. A naive approach is to deny (only) responses to queries for $X_a|X_B$. However, this rule alone will not prevent intruders from reconstructing the unreleased regression from other, releasable regressions. For example, suppose that the server provides regression coefficients for any query involving simple regressions

with X_{aB}. For $X_i, X_j \in X_{aB}$, an intruder can solve for the cross-product $\mathbf{S}_{aB}[i, j]$ by using the variance of the predictor, say X_j, and the released coefficient $b_{i|j}$ of X_j in the regression of X_i on X_j:

$$\mathbf{S}_{aB}[i, j] = b_{i|j}\, \mathbf{S}_{aB}[j, j]. \tag{1}$$

By fitting simple regressions for all pairs of variables, all terms in $\mathbf{S}_{aB}$ are determined, so that an intruder can reconstruct $X_a|X_B$ exactly.

More generally, any m unknown off-diagonal elements of $\mathbf{S}_{aB}$ can be reproduced exactly as long as the collection of released coefficients contains a system of m independent equations in these unknowns. Clearly, the coefficients for all simple regressions involving X_{aB} constitute such a collection. Other examples include coefficients for the set of all regressions of size k for any k (all $\mathbf{b}_{a|C}$ where $C \subset B$ and $|C| = k$) and coefficients for the set of sequential regressions, $\{\mathbf{b}_{a|i}, \mathbf{b}_{a|i,j}, \ldots, \mathbf{b}_{a|B}\}$, where $B = \{i, j, \ldots\}$. Thus, servers that release any regression as long as it has at least k predictors, or servers that release only one regression for each predictor size, do not protect $\mathbf{S}_{aB}$.

Reconstruction of some unreleased $X_a|X_B$ is not possible when at least one of the cross-products in $\mathbf{S}_{aB}$ cannot be determined from released information. To prevent some cross-product $\mathbf{S}_{aB}[i, j]$ from being reproduced exactly, the server must deny responses to queries involving X_i and X_j simultaneously. That is, the server cannot provide output for any query involving one of these variables as the outcome and the other as a predictor, or any query where both variables are predictors.

Although limiting the releases can prevent exact reconstruction of $\mathbf{S}_{aB}$, it still may be possible to bound closely the unknown elements of $\mathbf{S}_{aB}$. We next describe a procedure for finding upper and lower bounds for the unknown elements by exploiting the fact that $\mathbf{S}_{aB}$ is positive definite (denoted by $\mathbf{S}_{aB} \succ 0$).

Let $\mathcal{K} = \{(i, j) : \mathbf{S}_{aB}[i, j] \text{ is known}\}$ be the set of indices of the known elements of $\mathbf{S}_{aB}$. For each $(i, j) \in \mathcal{K}$, let s_{ij} denote the the value of its corresponding $\mathbf{S}_{aB}[i, j]$. For any $(l, m) \notin \mathcal{K}$, we can find the upper bound for $\mathbf{S}_{aB}[l, m]$ by solving the following optimization problem:

$$\begin{aligned} &\max \mathbf{S}_{aB}[l, m] \\ &\text{s.t.} \quad \begin{cases} \mathbf{S}_{aB}[i, j] = s_{ij} \quad \text{for all } (i, j) \in \mathcal{K} \\ \mathbf{S}_{aB} \succ 0. \end{cases} \end{aligned} \tag{2}$$

Define $\mathbf{F}_{pq}$, a matrix with the same dimensions as $\mathbf{S}_{aB}$, as follows. If $p = q$, then $\mathbf{F}_{pq}[i, j] = 1$ for $[i, j] = [p, q] (= [q, p])$ and is zero otherwise. If $p \neq q$, then $\mathbf{F}_{pq}[i, j] = 1/2$ for $[i, j] = [p, q]$ and $[i, j] = [q, p]$ and is zero otherwise. Then we can reformulate the optimization problem (2) as:

$$\begin{aligned} &\max \operatorname{Tr}(\mathbf{F}_{lm}\mathbf{S}_{aB}) \\ &\text{s.t.} \quad \begin{cases} \operatorname{Tr}(\mathbf{F}_{ij}\mathbf{S}_{aB}) = s_{ij} \quad \text{for all } (i, j) \in \mathcal{K} \\ \mathbf{S}_{aB} \succ 0, \end{cases} \end{aligned} \tag{3}$$

which is a Semidefinite Programming (SDP) problem expressed in standard form (Todd, 2001). Efficient algorithms and software implementations for SDP problems are available (Vandenberghe and Boyd, 1996; Todd, 2001). The lower bound for $\mathbf{S}_{aB}[l, m]$ is also obtained by solving the corresponding minimization problem.

These bounds provide the feasible range of values that each individual unknown element can take. When more than one element in $\mathbf{S}_{aB}$ is unknown, the individual feasible ranges determine a bounding box for the joint feasible region. It is possible to sample values of $\mathbf{S}_{aB}$ from the joint feasible region by sampling uniformly from the bounding box and then accepting or rejecting the sample point depending on whether

the resulting $\mathbf{S}_{aB}$ is positive definite. These values of $\mathbf{S}_{aB}$ in turn provide draws of feasible values of $\mathbf{b}_{a|B}$. When an intruder can obtain sufficiently tight bounds on $\mathbf{b}_{a|B}$, or on particular sensitive components of $\mathbf{b}_{a|B}$, inferential disclosures may occur.

Variants of this approach to obtaining bounds for unreleased coefficients can be applied to obtain approximate bounds in other models. Ordered categorical and dichotomous outcomes can be treated as continuous for purposes of using (2) and (3). Nominal variables with more than two categories can be split into a series of dichotomous indicator variables, which are then used in (2) and (3). Obtaining more precise bounds for other models is a subject for future research.

5 Disclosure Risk and Utility Measures for Model Servers

As for other SDL strategies, in the model server context it is essential to use quantitative measures of risk and utility to decide what is ultimately released. This section describes such measures generally and, as an entry point to a much larger research effort, presents specific instances for a linear regression setting.

In both cases, as well as in other settings such as table servers (Dobra et al., 2003, 2002), the distinction between risk and utility can be obscure. This is the heart of the risk-utility tradeoff problem: legitimate users and intruders may want the same, or nearly the same, things from the data.

5.1 General Measures

As suggested in §4, identity disclosure risk can be reduced by refusing to provide output for queries involving suspicious transformations. Hence for analysis servers we focus on measures of disclosure risk that reflect intruders' capability to predict accurately such values of individual units' attributes or relationships among sensitive attributes. We propose two broad classes of such risks. *In-sample prediction risk* refers to intruders' ability to predict accurately sensitive information for units in the database. An example is predicting an outcome for an atypical unit whose residual is small in some released or unreleased regression. *Out-of-sample prediction risk*, by contrast, refers to intruders' capability to predict closely sensitive information for units not in the database. An example is learning, either exactly or with little uncertainty, the values of the coefficients of a regression for a sensitive outcome, which can then be used to predict that sensitive variable for units not in the database.

Measures of utility quantify the amount of information contained in the answer space $\mathcal{A}$ relative to the information when no restrictions are made on the answer space. We propose two classes. *Volume* refers to the size of $\mathcal{A}$, for example the number of regression models in $\mathcal{A}$. *Statistical usefulness* refers to the extent to which the released information is useful for statistical inference. An example is the predictive accuracy of the models in $\mathcal{A}$. High statistical usefulness is not necessarily equivalent to large volume: a small answer space may well contain higher quality models than some larger one. Utility also can incorporate domain knowledge: for instance, to satisfy users' needs, agencies may decide particular relationships must be released.

These classes of risk and utility measures are related to the predictive accuracy of the models in $\mathcal{A}$. Risk and utility do have a distinction in our formulation: utility is always calculated using the information in released models, whereas risk can be calculated using what is inferred about unreleased models.

5.2 Risk and Utility Measures for a Linear Regression Setting

Risk and utility measures obviously depend on the types of models in the query space $\mathcal{Q}$. To make the ideas concrete and to illustrate the general notions of risk and utility, suppose that $\mathcal{Q}$ corresponds to linear regressions, and that the database $\mathcal{D}$ contains a single sensitive variable that the agency does not want intruders to be able to predict too accurately from released regressions on the other variables in $\mathcal{D}$.

Using the notation of §4.2, let X_0 be the sensitive variable, and let $X_1, \ldots, X_d$ be the other variables. We assume the agency is using a static model server, and thus seeks to determine an optimal answer space that results in high data utility with acceptable disclosure risk. For simplicity, we assume that no transformations of the X_i are allowed.

For this $\mathcal{Q}$, there are 2^d queries involving X_0 as the dependent variable, corresponding to 2^{2^d} possible choices for $\mathcal{A}$. Calculating the risk and utility of all these is infeasible even for small values of d. Therefore, we restrict $\mathcal{A}$ to a more manageable subset, which we call $\mathcal{A}_{\text{supp}}$ ("supp" denotes suppressed variables), defined as follows.

Suppose that $X_{\text{supp}} \subseteq \{X_1, X_2, \ldots, X_d\}$ and $X_{\text{free}} = \{X_1, X_2, \ldots, X_d\} \setminus X_{\text{supp}}$, and let $\mathcal{A}_{\text{supp}}$ be the answer space containing all regressions *except* those with X_0 as the response and at least one of the variables in X_{supp} as a predictor, or vice versa. In this case, any user, legitimate or not, can determine exactly all cross-products between X_0 and the variables of X_{free}, between the variables of X_{supp} and those of X_{free}, and among the variables of X_{free}.

Intruders may attempt to use $\mathcal{A}_{\text{supp}}$ to reproduce any of the cross-products involving X_0 and elements of X_{supp}, and hence any of the associated regression coefficients. Note that the strategy of predicting attributes in X_{supp} from attributes X_{free} is not effective: the information needed to do this is already available in the ellipsoid obtained using Result 1 (see the Appendix).

In addition to restricting the search space, using $\mathcal{A}_{\text{supp}}$ has practical benefits. Any regression that does not involve X_0 can be fit, which increases both volume and statistical usefulness. Relationships among predictors of X_0 can be examined, which increases statistical usefulness by facilitating checks for multicollinearity.

We search for an optimal $\mathcal{A}$ over possible specifications of X_{supp} and corresponding $\mathcal{A}_{\text{supp}}$. Our specific risk and utility measures are based on users' ability to predict the unknown cross-products between X_0 and X_{supp} using output from $\mathcal{A}_{\text{supp}}$. These entries, as well as the rest of the cross-products matrix $\mathbf{S}$, can be partitioned as

$$\mathbf{S} = \left[\begin{array}{c|cc} s_{00} & \mathbf{s}_{\text{supp}}^t & \mathbf{s}_{\text{free}}^t \\ \hline \mathbf{s}_{\text{supp}} & \multicolumn{2}{c}{\multirow{2}{*}{$\mathbf{S}_D$}} \\ \mathbf{s}_{\text{free}} & & \end{array}\right], \tag{4}$$

where $D = \{X_1, \ldots, X_d\}$, $\mathbf{s}_{\text{supp}}$ is the cross-products between X_0 and X_{supp}, and $\mathbf{s}_{\text{free}}$ contains the cross-products between X_0 and X_{free}.

When all elements of $\mathbf{S}$ are known except for the strip $\mathbf{s}_{\text{supp}}$, the feasible values of $\mathbf{s}_{\text{supp}}$ must lie in the interior of an ellipsoid, as shown in Result 1 in the Appendix. We use this ellipsoid to construct specific measures of in-sample and out-of-sample disclosure risk.

Specific Risk Measures. *Residual risk* R_{res} quantifies users' ability to predict X_0 for particular subsets of units in the data, for example those with atypical attribute values. The risk measure is the reciprocal of the square root of the average of the squared residuals for the selected subset, obtained from the regression $X_0|X_{\text{free}}$.

Prediction risk R_{pred} quantifies users' ability to predict X_0 from the largest possible regression, namely $X_0|X_1, X_2, \ldots, X_d$. When some regressions are suppressed, i.e., $X_{\text{supp}} \neq \emptyset$, the user draws feasible values

of the unreleased $\mathbf{s}_{\text{supp}}$ from the ellipsoid as described in §4, thereby generating feasible coefficients for $X_0|X_1, X_2, \ldots, X_d$. The risk measure is the average value of R^2 for these feasible regressions, which summarizes the predictive ability of the feasible models. When drawing $\mathbf{s}_{\text{supp}}$ from the ellipsoid uniformly, the sampling distribution of the average R^2 of the feasible models can be determined analytically, as is shown in Result 2 in the appendix.

These measures can be adjusted to meet particular needs. For R_{pred}, values of $\mathbf{s}_{\text{supp}}$ can be drawn non-uniformly, for example, to reflect domain knowledge by giving more weight to feasible regions consistent with estimated coefficients available from published analyses. The values can be drawn so that certain coefficients are always positive or always negative. Rather than the average of the feasible R^2, the measure can be some function of the bounds on the predicted values of X_0 implied by the feasible regressions. Similarly, for R_{res}, the residuals can be based on the feasible regressions rather than the released ones. Or, the measure could be based on the relative, absolute residuals rather than squared residuals.

Specific Utility Measures. To measure volume, we use the dimension of X_{free}. For statistical usefulness, we present two measures. *Unweighted accuracy* U_{rsq} is the R^2 of $X_0|X_{\text{free}}$. *Weighted accuracy* U_{rsqwt} adds weights w_i that reflect the importance of the variables: $U_{\text{rsq}} + \sum_{i \in \text{free}} w_i$, allowing agencies to incorporate domain knowledge into utility measures. Each w_i can be interpreted as the "R^2 points" gained by including X_i in X_{free}. Setting $w_i = 1$ forces X_i to be in X_{free}. Setting all $w_i = 0$ corresponds to having no domain knowledge-based preferences about which variables are included.

Other utility measures targeted at estimation rather than prediction can be devised, and are associated closely with the bounds derived in the Appendix.

5.3 Illustrating the Measures: A Simulation Study

We now illustrate the risk and utility measures R_{pred}, R_{res}, U_{rsq} and U_{rsqwt} using two simulated databases. Both comprise 200 records, and contain one response variable X_0 and nine predictors $X_1, X_2, \ldots, X_9$. In Data Set I, X_1, X_2, X_3 are highly correlated, and each is highly correlated with X_0. Data Set II has no strong relationships among the variables.

For R_{res}, we select the units with the highest 5% of the X_0 values as the target set, so that the agency is protecting units with extreme values of X_0. For R_{pred}, we draw feasible values of $\mathbf{s}_{\text{supp}}$ uniformly from the ellipse. For U_{rsqwt}, we set the w_i to equal the R^2 of the simple linear regression of X_0 on X_i. In reality, for R_{res} the agency specifies the target set, and for U_{rsqwt} the agency specifies weights based on domain knowledge.

The measures are evaluated on each of 510 possible releases; the two unevaluated regressions include X_0 on the intercept only and $X_0|X_1, \ldots, X_9$. Figures 1–3 display scatterplots of the utility measures versus the risk measures. Each point represents the value of the utility and risk functions for a particular candidate release, $\mathcal{A}_{\text{supp}}$. These displays can be used to select $\mathcal{A}$. In all figures, color indicates the dimension of X_{free}.

Behavior of the Risk and Utility Measures. For Data Set I, Figure 1 shows that, as expected, utility generally increases as risk increases. The precise relationship depends on the risk-utility combination, suggesting that these measures capture different aspects of risk and utility for this data set. This results from the structure of Data Set I: any release containing one or more of (X_1, X_2, X_3) has high risk and utility, and any other release has low risk and utility. In Figure 2, the colored points (those releases for which X_{free} does not contain any of X_1, X_2, X_3) all lie in the low-risk, low-utility region. The effect of (X_1, X_2, X_3) also explains why no clear dimension effect is evident in Figure 1.

For Data Set II, Figure 3 indicates a clear dimension effect. This is because no predictors are strong, so that increasing the number of predictors raises all measures of risk and utility.

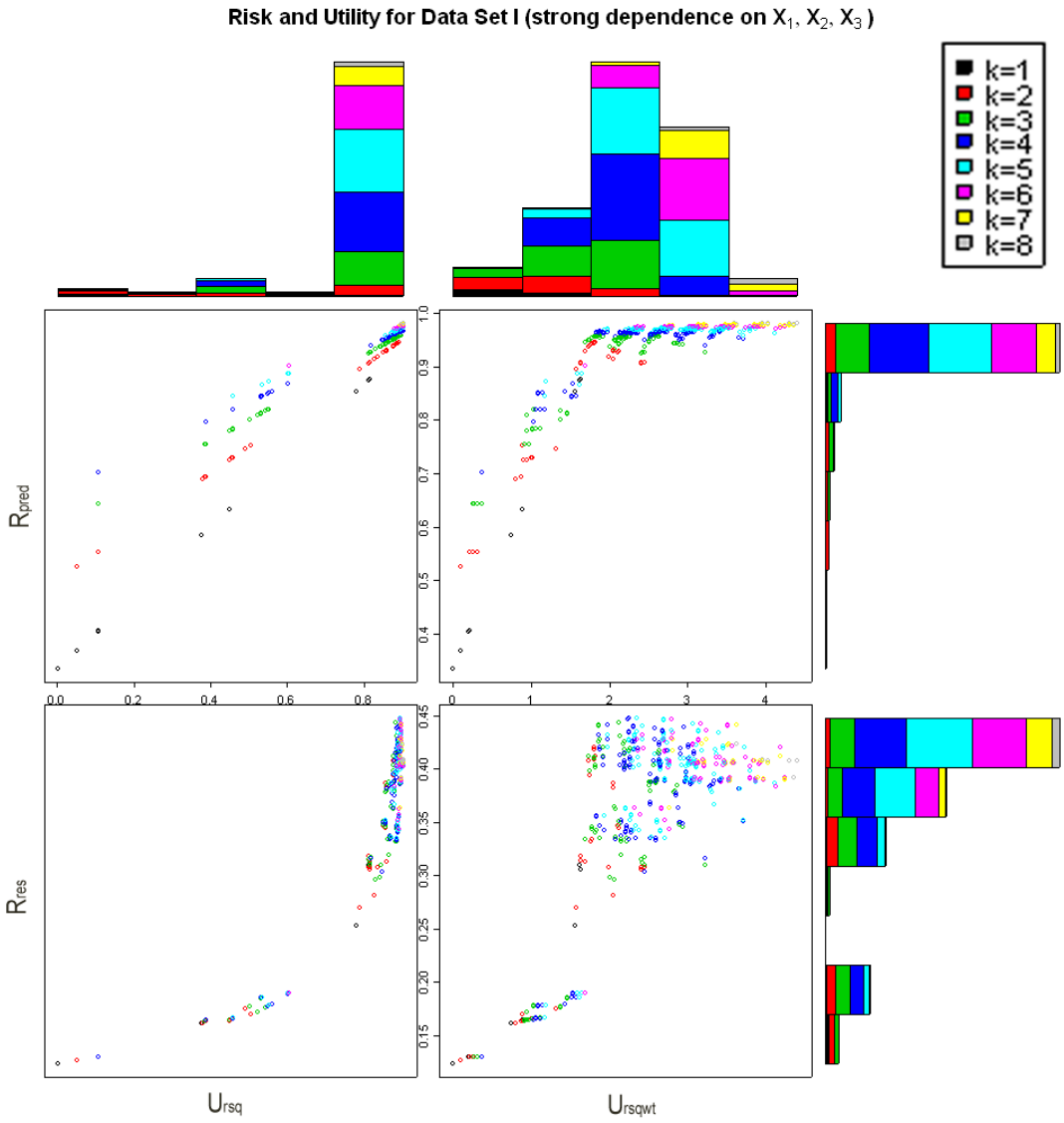

Figure 1: Risk-utility scatterplots for both risk and utility measures and corresponding univariate histograms for Data Set I.

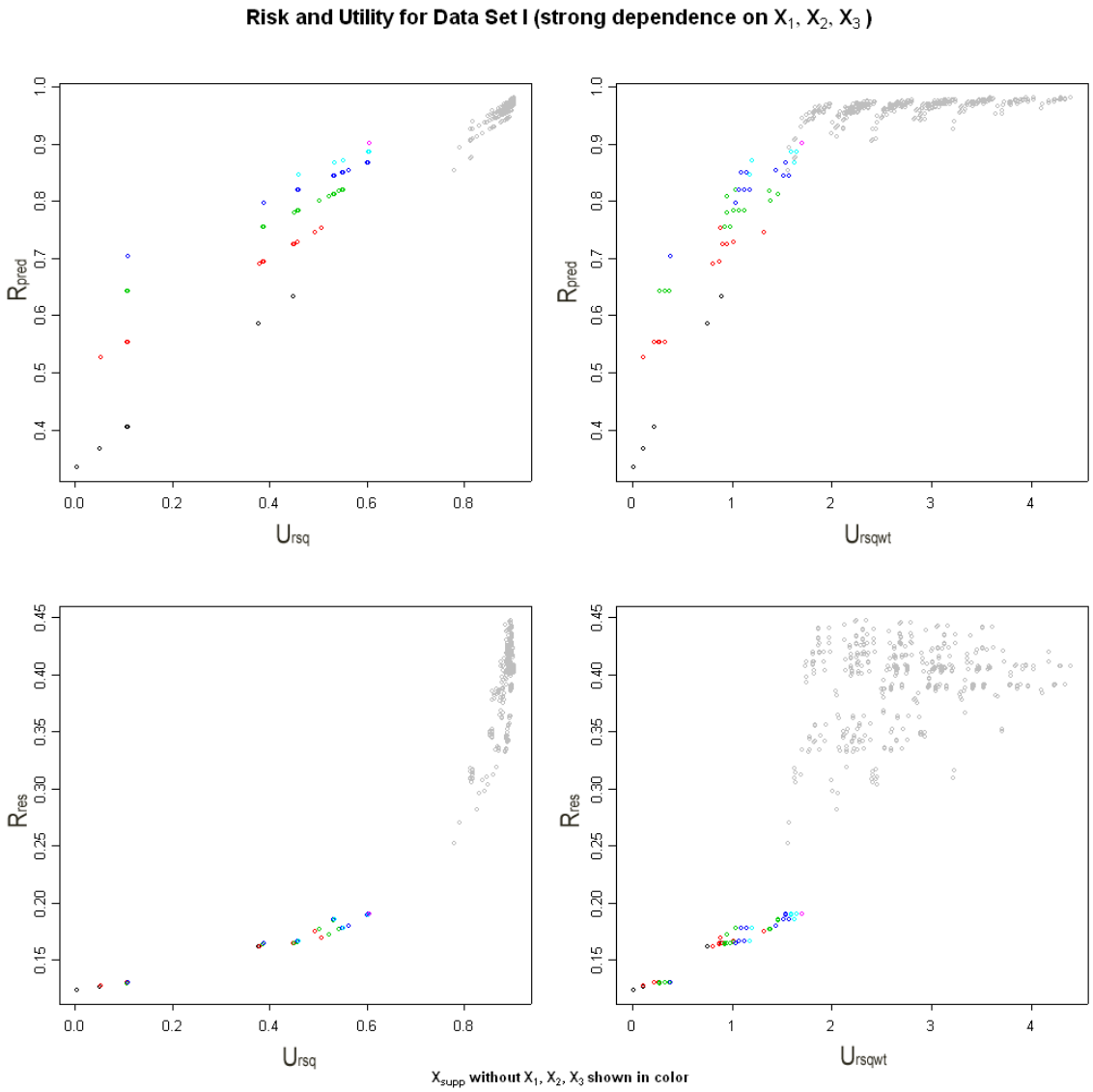

Figure 2: Risk-utility scatterplots for both risk and utility measures for Data Set I. Colored points are those releases in which $\{X_1, X_2, X_3\} \notin \mathbf{X}_{\text{free}}$.

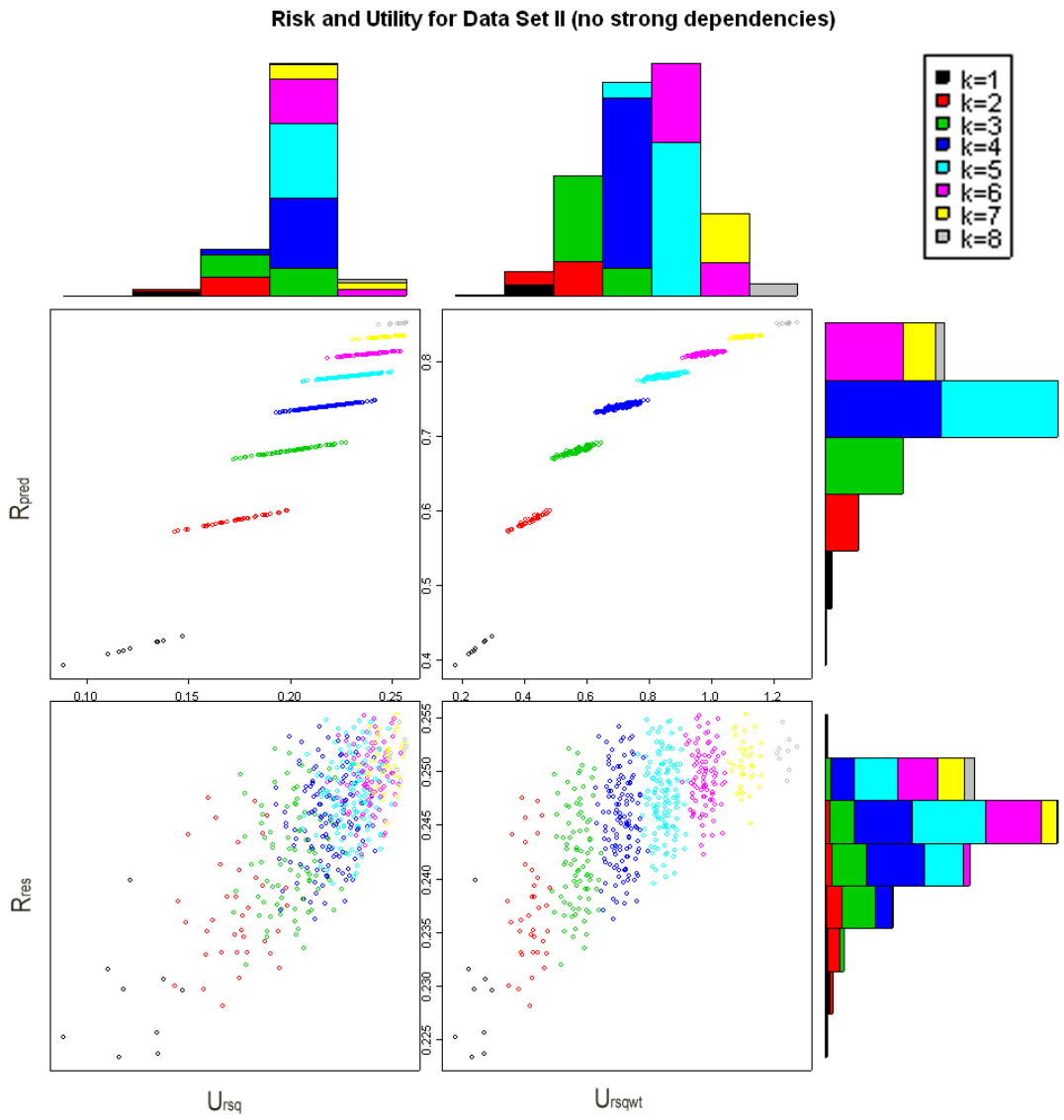

Figure 3: Risk-utility scatterplots for both risk and utility measures and corresponding univariate histograms for Data Set II.

Selecting an Optimal Release. We now illustrate how $\mathcal{A}_{\text{supp}}$ can be determined from the risk-utility plot of the candidate releases. As mentioned in §3, general approaches include optimization of $U(\mathcal{A})$ subject to an upper threshold on $R(\mathcal{A})$ and selection of $\mathcal{A}_{\text{supp}}$ from a frontier of undominated candidate spaces—those for which no other candidate release has both lower disclosure risk and higher data utility. These approaches are displayed in Figure 4.

To illustrate the risk threshold approach, suppose the agency seeks to prevent intruders from predicting X_0 for the chosen target units within 5% on average, which corresponds roughly to a R_{res} threshold of 0.2. Based on this, we pick the release candidate with highest U_{rsq} and $R_{\text{res}} < 0.2$, yielding as the optimal release $\mathcal{A}_{\text{supp}}$ associated with $X_{\text{supp}} = \{X_1\}$, so that $X_{\text{free}} = \{X_2, X_3, \ldots, X_9\}$.

To illustrate the frontier approach, the agency first defines a function of risk and utility that quantifies the "benefit" to the agency for specified values of risk and utility. Contours of this function on the risk-utility plane show how the agency is willing to trade risk for utility for a fixed level of "benefit". The agency then finds the point on the curve connecting all undominated release candidates—the frontier shown in color in the right panel of Figure 4—that is the first to touch a risk-utility trade-off contour of highest "benefit." In the figure, the trade-off contours are linear; "benefit" increases as the line is shifted in a southeast direction. The line is moved northwest, with the slope kept constant, until it touches a point on the frontier, and this point corresponds to the optimal release. In Data Set I, this procedure again picks $\mathcal{A}_{\text{supp}}$ defined by $X_{\text{supp}} = \{X_1\}$ as the optimal release.

Clearly risk-utility plots and optimal releases based on them will vary for different data sets. For instance, Data Set II would yield very different risk-utility values for the optimal release.

6 Discussion

Much of our discussion of disclosure risk and data utility in model servers has been in the context of linear regression, and our illustrative example involved protecting a single variable. Protecting multiple variables and dealing with models other than linear regressions complicate the measurement of risk and utility. We document here some of these additional challenges and suggest paths for future research.

In some databases, relationships involving multiple variables are subject to inferential disclosure. Release decisions for individual variables necessarily interact and can affect risk and utility. For example, suppose X_a and X_b are sensitive and can be predicted closely using other variables. Protecting X_a by prohibiting a set of variables from appearing in models with X_a also restricts the answer model space for X_b, and *vice versa*.

For a small number of variables, it is possible to enumerate all regressions using the sensitive variables as outcomes, and to compute the risks and utilities for each possible release. This approach is computationally challenging for data sets with many variables. It may be possible to consider only a small set of predictors as candidates for those that may not appear with the sensitive variables. Developing effective search strategies, as well as measures of combined risk and utility, is an area for further research.

It is important from a utility perspective to provide output for models involving transformations $g(X)$ that do not result in disclosures (see §4). Operationally, such transformations are not a problem for the server: the user can submit and receive output from models with $g(X)$ replacing X, and agencies can protect relationships involving $g(X)$ rather than X. However, release strategies designed to protect relationships involving $g(X)$ do not necessarily protect relationships involving other transformations of X. Clear-cut and universally acceptable transformations of the data can be implemented prior to the agency's releasing the data. Beyond that, one approach is to disallow any transformations of the data, but at a high cost in data utility. A less restricted alternative is to limit the space of permissible transformations (for example, to

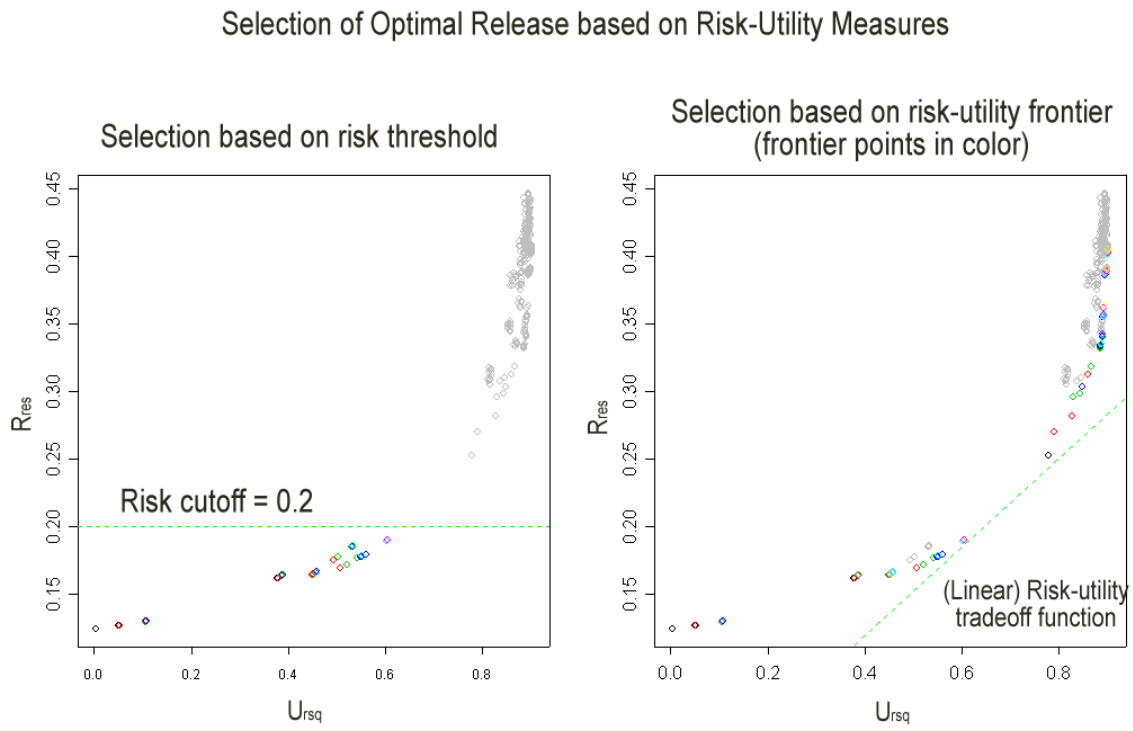

Figure 4: Risk-utility plots for selecting an optimal release. *Left:* release based on a risk threshold. *Right:* release based on a risk-utility frontier.

logarithms, squares and square roots) or to limit the models that can be fit with them (for example, whenever $g(X_a)$ and $g(X_b)$ are not allowed to appear simultaneously in models, all other transformations of X_a and X_b are prohibited as well). Finding methods of assessing disclosure risks that account for transformations, even when the space of transformations is restricted, is an extremely challenging problem for further research.

As mentioned in §4.2, approximate bounds for unreleased regression coefficients of complex models—such as generalized linear models or generalized additive models—can be obtained by approximating the complex model with a linear regression. It may be possible to obtain sharper bounds on estimated coefficients. For example methods for bounding cells of tabular data correspond to bounds on coefficients of particular log-linear models (Dobra et al., 2002, 2003). Research on bounds for complex models would have useful applications for data dissemination even outside the model server context.

It is prudent for agencies to use relevant domain knowledge when deciding what can be released by a server. As touched on in §5, such considerations can be incorporated into the risk and utility measures. Examples incorporating domain knowledge for genuine data would be useful blueprints for agencies considering the analysis server approach.

Acknowledgements

This research was supported by NSF grant IIS–0131884 to the National Institute of Statistical Sciences.

References

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

A. Dobra, A. F. Karr, S. E. Fienberg, and A. P. Sanil. Software systems for tabular data releases. *Int. J. Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):529–544, 2002.

A. Dobra, A. F. Karr, and A. P. Sanil. Preserving confidentiality of high-dimensional tabulated data: Statistical and computing issues. *Statistics and Computing*, 13:363–370, 2003.

J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing SDC methods for microdata on the basis of information loss and disclosure risk. *Presented at UNECE Workshop on Statistical Data Editing*, May, 2001.

G. T. Duncan, V. A. de Wolf, T. B. Jabine, and M. L. Straf. Report of the Panel on Confidentiality and Data Access. *J. Official Statist.*, 9:271–274, 1993.

G. T. Duncan, S. A. Keller-McNulty, and S. L. Stokes. Disclosure risk vs. data utility: The R-U confidentiality map. *Submitted to Manag. Sci.*, 2002.

G. T. Duncan and D. Lambert. Disclosure-limited data dissemination (with discussion). *J. Amer. Statist. Assoc.*, 81:10–28, 1986.

G. T. Duncan and S. Mukherjee. Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. *Journal of the American Statistical Association*, 95:720–729, 2000.

Federal Committee on Statistical Methodology. *Report on Statistical Disclosure Limitation Methodology*. US Office of Management and Budget, Washington, 1994.

S. E. Fienberg, U. E. Makov, and A. P. Sanil. A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *J. Official Statist.*, 13:75–89, 1997.

S. E. Fienberg, U. E. Makov, and R. J. Steele. Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14:485–502, 1998.

S. E. Fienberg, R. J. Steele, and U. E. Makov. Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and log-linear models. In *Proceedings of Bureau of Census 1996 Annual Research Conference*, pages 87–105, 1996.

W. A. Fuller. Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9: 383–406, 1993.

S. Gomatam, A. F. Karr, and A. P. Sanil. Data swapping as a decision problem. *J. Official Statist.*, 2003. Submitted for publication.

I. Guttman. *Linear Models: An Introduction*. Wiley, 1982.

A. F. Karr, A. Dobra, and A. P. Sanil. Table servers protect confidentiality in tabular data releases. *Comm. ACM*, 46(1):57–58, 2003.

A. F. Karr, X. Lin, J. P. Reiter, and A. P. Sanil. Analysis of integrated data without data integration. *Chance*, 17(3):26–29, 2004a.

A. F. Karr, X. Lin, J. P. Reiter, and A. P. Sanil. Secure regression on distributed databases. *J. Computational and Graphical Statist.*, 2004b. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

S. Keller-McNulty and E. A. Unger. A database system prototype for remote access to information based on confidential data. *Journal of Official Statistics*, 14:347–360, 1998.

D. Lambert. Measures of disclosure risk and harm. *J. Official Statist.*, 9(2):313–331, 1993.

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

G. Paass. Disclosure risk and disclosure avoidance for microdata. *J. Business and Economic Statis.*, 6: 487–500, 1988.

M. A. Palley and J. S. Simonoff. The use of regression methodology for the compromise of confidential information in statistical databases. *ACM Transacations on Database Systems*, 12:593–608, 1987.

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

J. P. Reiter. Satisfying disclosure restrictions with synthetic data sets. *J. Official Statist.*, 18:531–544, 2002.

J. P. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:81–88, 2003a.

J. P. Reiter. Model diagnostics for remote access regression servers. *Statistics and Computing*, 13:371–380, 2003b.

J. P. Reiter. Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *J. Royal Statist. Soc., Series A*, 2004. To appear.

J. P. Reiter and C. Kohnen. Categorical data regression diagnostics for remote access servers. *J. Statistical Computation and Simulation*, 2004. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

S. Rowland. An examination of monitored, remote access microdata access systems. In *National Academy of Sciences Workshop on Data Access*, 2003.

D. B. Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:462–468, 1993.

A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter. Privacy preserving analysis of vertically partitioned data using secure matrix products. *J. Official Statist.*, 2004a. Submitted for publication. Available on-line at www.niss.org/dgii/technicalreports.html.

A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter. Privacy preserving regression modelling via distributed computation. In *Proc. Tenth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, 2004b. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

B. Schouten and M. Cigrang. Remote access systems for statistical analysis of microdata. *Statistics and Computing*, 13:381–389, 2003.

L. Sweeney. Computational disclosure control for medical microdata: the Datafly system. In *Proceedings of an International Workshop and Exposition*, pages 442–453, 1997.

M. J. Todd. Semidefinite optimization. *Acta Numerica*, 10:515–560, 2001.

L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

L. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*. Springer–Verlag, New York, 2001.

L. C. R. J. Willenborg and T. de Waal. *Statistical Disclosure Control in Practice*. Springer–Verlag, New York, 1996.

W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata. *Inf. Control in Statist. Databases*, 2002.

Appendix

Result 1 *Positive definiteness of* $\mathbf{S}$ *ensures that* $\mathbf{s}_{\text{supp}}$ *lies within the ellipsoid defined by*

$$(\mathbf{s}_{\text{supp}} - \mathbf{c})^t \mathbf{B} (\mathbf{s}_{\text{supp}} - \mathbf{c}) < 1, \tag{5}$$

where $\mathbf{B} = \left(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\right)^{-1} / r$*, with* $\mathbf{S}_{11}, \mathbf{S}_{12}, \mathbf{S}_{21}, \mathbf{S}_{22}$ *being the appropriately partitioned elements of* $\mathbf{S}_D$ *(with partitions corresponding to the lengths of* $\mathbf{s}_{\text{supp}}$ *and* $\mathbf{s}_{\text{free}}$ *respectively),* $r = s_{00} - \mathbf{s}_{\text{free}}^t \mathbf{S}_{22}^{-t} \mathbf{s}_{\text{free}}$*, and* $\mathbf{c} = \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{s}_{\text{free}}$ *is the center of the ellipsoid.*

Proof In the partition of $\mathbf{S}$ given in (4), let $\mathbf{S}_D$ be partitioned as follows:

$$\mathbf{S}_D = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix},$$

where $\mathbf{S}_{11}$ has dimension $k \times k$, and $\mathbf{S}_{22}$ has dimension $(d-k) \times (d-k)$. Let $\mathbf{S}_D^{-1}$ have the corresponding partition:

$$\mathbf{S}_D^{-1} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

(Since $\mathbf{S}_D$ is positive definite, its inverse exists.)

As $\mathbf{S}$ is also positive-definite,

$$\det(\mathbf{S}) = \det(\mathbf{S}_D)\det(s_{00} - \mathbf{s}^t\mathbf{S}_D^{-1}\mathbf{s}),$$

where $\mathbf{s}^t = (\mathbf{s}_{\text{supp}}^t \;\; \mathbf{s}_{\text{free}}^t)$. Since $\mathbf{S}$ and $\mathbf{S}_D$ are both strictly positive definite, $s_{00} - \mathbf{s}^t\mathbf{S}_D^{-1}\mathbf{s} > 0$, so that

$$\frac{\mathbf{s}^t\mathbf{S}_D^{-1}\mathbf{s}}{s_{00}} < 1 \tag{6}$$

and therefore

$$\frac{\mathbf{s}_{\text{free}}^t\mathbf{A}_{22}\mathbf{s}_{\text{free}} + 2\mathbf{s}_{\text{free}}^t\mathbf{A}_{21}\mathbf{s}_{\text{supp}} + \mathbf{s}_{\text{supp}}^t\mathbf{A}_{11}\mathbf{s}_{\text{supp}}}{s_{00}} < 1. \tag{7}$$

With $\mathbf{c} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{s}_{\text{free}}$, (7) can be rewritten as:

$$\frac{\mathbf{s}_{\text{free}}^t(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})\mathbf{s}_{\text{free}} + (\mathbf{s}_{\text{supp}} - \mathbf{c})^t\mathbf{A}_{11}(\mathbf{s}_{\text{supp}} - \mathbf{c})}{s_{00}} < 1. \tag{8}$$

This can be further rewritten as $(\mathbf{s}_{\text{supp}} - \mathbf{c})^t\mathbf{A}_{11}(\mathbf{s}_{\text{supp}} - \mathbf{c}) < r$, where $r = s_{00} - \mathbf{s}_{\text{free}}^t(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})\mathbf{s}_{\text{free}}$. That is,

$$(\mathbf{s}_{\text{supp}} - \mathbf{c})^t\mathbf{B}(\mathbf{s}_{\text{supp}} - \mathbf{c}) < 1, \tag{9}$$

where $\mathbf{B} = \mathbf{A}_{11}/r$. As $\mathbf{A}_{11}$ is strictly positive definite, and $r > 0$, the inequality in (9) represents the interior of an ellipsoid.

Using expressions for inverses of partitioned matrices (Guttman, 1982), we can rewrite $\mathbf{c}$ and r in terms of elements of $\mathbf{S}$ as $\mathbf{c} = \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{s}_{free}$ and

$$r = s_{00} - \mathbf{s}_{\text{free}}^t\mathbf{S}_{22}^{-1}\mathbf{s}_{\text{free}}. \tag{10}$$

Note 1 The volume of the ellipsoid is given by

$$V_{\mathcal{E}_k} = \frac{\pi^{\frac{k}{2}}}{\Gamma(1+\frac{k}{2})} \cdot \frac{1}{\sqrt{\det(\mathbf{B})}} = \frac{\pi^{\frac{k}{2}}}{\Gamma(1+\frac{k}{2})} \cdot \frac{(s_{00} - \mathbf{s}_{\text{free}}^t\mathbf{S}_{22}^{-1}\mathbf{s}_{\text{free}})^{k/2}}{\sqrt{\det(\mathbf{A}_{11})}}, \tag{11}$$

where $\Gamma(\cdot)$ is the gamma function. Note that (11) can be re-expressed as $V_{\mathcal{E}_k} = V_k/\sqrt{\det(\mathbf{B})}$, where V_k is the volume of the k-dimensional unit hypersphere.

Note 2 For the regression of X_0 on $\{X_1, X_2, \ldots, X_d\}$, the coefficient of determination R^2 is $\hat{X}_0^t\hat{X}_0/X_0^tX_0 = X_0^t\hat{X}_0/X_0^tX_0$, which in terms of elements of the S matrix is $\mathbf{s}^t\mathbf{S}_D^{-1}\mathbf{s}/s_{00}$. We can see from (6) that the interior of the ellipse defines the region (in $\mathbf{s}_{\text{supp}}$-space) where the R^2 is less than 1.

Note 3 Similarly, for the regression of X_0 on the variables in X_{free}, the coefficient of determination, in terms of elements of the S matrix, is $\mathbf{s}_{\text{free}}^t \mathbf{S}_{22}^{-1} \mathbf{s}_{\text{free}} / s_{00}$. We will denote this quantity by ρ.

Note 4 From (6)–(8), R^2 can be written as

$$\begin{aligned} R^2 &= \frac{\mathbf{s}_{\text{free}}^t (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})\mathbf{s}_{\text{free}} + (\mathbf{s}_{\text{supp}} - \mathbf{c})^t \mathbf{A}_{11} (\mathbf{s}_{\text{supp}} - \mathbf{c})}{s_{00}} \\ &= \frac{\mathbf{s}_{\text{free}}^t \mathbf{S}_{22}^{-1} \mathbf{s}_{\text{free}} + (\mathbf{s}_{\text{supp}} - \mathbf{c})^t \mathbf{A}_{11} (\mathbf{s}_{\text{supp}} - \mathbf{c})}{s_{00}}. \end{aligned} \tag{12}$$

Since both $\mathbf{S}_{22}^{-1}$ and $\mathbf{A}_{11}$ are positive definite, and $\mathbf{s}_{\text{free}}$ is known, R^2 is minimized for $\mathbf{s}_{\text{supp}} = \mathbf{c}$, so that

$$R^2_{\text{min}} = \frac{\mathbf{s}_{\text{free}}^t (\mathbf{S}_{22}^{-1}) \mathbf{s}_{\text{free}}}{s_{00}} = \rho.$$

Hence the ellipse that defines the feasible region for $\mathbf{s}_{\text{supp}}$ corresponds to $\rho \leq R^2 < 1$.

Result 2 *If $\mathbf{s}_{\text{supp}}$ is distributed uniformly over its support (given by the ellipsoid from (9)), then the distribution of R^2, the coefficient of determination for the regression of X_0 on $\{X_1, X_2, \ldots, X_d\}$, has density function*

$$f_{R^2}(u) = \frac{k/2}{(1-\rho)^{k/2}} (u - \rho)^{\frac{k}{2}-1} \qquad \text{for } \rho \leq u < 1 \tag{13}$$

and expectation

$$E(R^2) = \frac{k + 2\rho}{k+2}. \tag{14}$$

Proof We know that the defining condition for the ellipsoid is given by (6), which is equivalent to the requirement that $R^2 < 1$ (see Note 1). Moreover, $\rho \leq R^2$ since $\mathbf{s}_{\text{free}}$ is known (see Note 4). Let us denote k-dimensional ellipsoid in by $\mathcal{E}_k(1)$. We can also define the k-dimensional ellipsoid that corresponds to the condition $R^2 < u$ (with $\rho \leq u$) by $\mathcal{E}_k(u)$. Analogous to the derivation of Result 1, $\mathcal{E}_k(u)$ is defined by:

$$(\mathbf{s}_{\text{supp}} - \mathbf{c})^t \mathbf{B}_u (\mathbf{s}_{\text{supp}} - \mathbf{c}) < 1, \tag{15}$$

where $\mathbf{B}_u = \mathbf{A}_{11}/r_u$, with $\mathbf{A}_{11}$ as in Result 1 and

$$r_u = u s_{00} - \mathbf{s}_{\text{free}}^t \mathbf{S}_{22}^{-1} \mathbf{s}_{\text{free}} \tag{16}$$

If $\mathbf{s}_{\text{supp}}$ is distributed uniformly over its support ($\mathcal{E}_k(1)$) then $F_{R^2}(u) = \Pr(R^2 \leq u)$ the ratio of the volumes of the two ellipsoids: $V(\mathcal{E}_k(u))/V(\mathcal{E}_k(1))$. From (11),

$$\begin{aligned} F_{R^2}(u) = \Pr(R^2 \leq u) = \frac{V(\mathcal{E}_k(u))}{V(\mathcal{E}_k(1))} &= \left(\frac{r_u}{r}\right)^{k/2} \\ &= \left(\frac{u s_{00} - \mathbf{s}_{\text{free}}^t \mathbf{S}_{22}^{-1} \mathbf{s}_{\text{free}}}{s_{00} - \mathbf{s}_{\text{free}}^t \mathbf{S}_{22}^{-1} \mathbf{s}_{\text{free}}}\right)^{k/2} \\ &= \left(\frac{u - \rho}{1 - \rho}\right)^{k/2}. \end{aligned} \tag{17}$$

Differentiation of (17) yields (13), and a straightforward expectation calculation yields (14).