Journal of Privacy and Confidentiality (2009) **1**, Number 2, pp. 197–211

Secure Statistical Analysis of Distributed Databases, Emphasizing What We Don't Know

Alan F. Karr*

Abstract. Over the past several years, the National Institute of Statistical Sciences (NISS) has developed methodology to perform statistical analyses that, in effect, integrate data in multiple, distributed databases, but without literally bringing the data together in one place. In this paper, we summarize that research, but focus on issues that are not understood. These include inability to perform exploratory analyses and visualizations, protections against dishonest participants, inequities between database owners and lack of measures of risk and utility.

Keywords: data confidentiality, distributed databases, secure multi-party computation

1 Introduction

Many government, industrial, and academic investigations require statistical analyses based on data stored in multiple distributed databases, often each with a different owner. But, barriers to the actual integration of the databases are numerous:

Confidentiality, as in "official statistics" (Karr et al., 2004, 2005b; Sanil et al., 2004, 2007), or homeland security (Karr et al., 2006b) settings.

Proprietary data, as in the chemical database example in §3.2.

Scale: Despite advances in networking technology, the only sure way to move a petabyte of data from one point today to another point tomorrow may be by using FedEx or UPS.

For many analyses (using techniques from computer science known generically as *secure multi-party computation*), the database owners can share sufficient statistics anonymously, but in a way that the analysis can be performed in a statistically valid manner. The protocols provide the owners protection from one another in the sense that while each owner can compare the global analysis to the same analysis on its own data, it is not able to attribute any characteristics of the discrepancies to other specific databases.

In this paper, "database" means a flat file in which rows represent data subjects and columns represent attributes. We term the database owners "agencies," although in the example in §3.2 they are companies. There are two structured data partitioning models. For *horizontally partitioned data*, the data subjects are partitioned among the

*National Institute of Statistical Sciences, Research Triangle Park, NC, mailto:karr@niss.org

databases containing the same attributes.[1] For *vertically partitioned data*, the attributes are partitioned among the databases. More complex partitions are discussed in §5.

This paper summarizes computational protocols, but focuses on *what is not understood*. §2 introduces secure multi-party computation (SMPC) and the protocol we use—secure summation. §3 presents protocols for a variety of analyses for horizontally partitioned data, including regression, contingency tables, and maximum likelihood for exponential families. Briefer discussion of vertically partitioned data appears in §4, along with a similarly brief discussion of complex data partitions in §5. In each of these sections, we discuss the numerous gaps in our knowledge, ranging from the conceptual to the computational. Conclusions appear in §6.

2 Secure Multi-Party Computation

Consider K agencies with values $v_1, \ldots, v_K$ who wish to evaluate a known function f at these values subject to four constraints:

C1: The correct value $f(v_1, \ldots, v_K)$ is obtained and known to all agencies.

C2: No agency j learns more about the other agencies' values $\mathcal{V}_{-j} = \{v_k : k \neq j\}$ than it can deduce from v_j and $f(v_1, \ldots, v_K)$.

C3: No trusted third party—human or machine—is part of the process.

C4: *Semi-honesty.* Agencies perform agreed-upon computations correctly using their true data. However, they are permitted to retain the results of intermediate computations.

The computer science literature contains many papers on the theory of SMPC; general references are Goldwasser (1997) and Yao (1982). There are many fewer implemented algorithms, let alone functioning software systems.

In §3 we employ *secure summation* (Benaloh, 1987): $f(v_1, \ldots, v_K) = v_1 + \cdots + v_k$, denoted by V. The steps are as follows:

Initialization: Agency 1 generates and retains a very large, complex random number R, adds R to its value v_1, and sends $R + v_1$ to agency 2.

Iteration: Agency 2 adds its value v_2 to $R + v_1$, sends the result to agency 3, and so on.

Sharing: Finally, agency 1 receives $R + v_1 + \cdots + v_K = R + V$ from agency K, subtracts R, and shares the result V with the other agencies.

There are issues with secure summation. First, it needs a "good" random number, in particular, one not ending in a string of zeroes, and not recoverable by guessing

[1] There are some subtleties associated with "the same;" see §3.4.

the seed. Second, collusion is possible: agencies $j-1$ and $j+1$, without sharing private information, can determine v_j.[2] Production-quality implementation is subtle: the process must be safe from outsiders, masqueraders, and (if there is one) a central server. Finally, secure summation is not a Nash equilibrium: it breaks if semi-honesty fails (Karr et al., 2007).

3 Secure Analysis of Horizontally Partitioned Data

The protocols described here are based on one underlying idea: if the analysis uses *sufficient statistics that are additive across agencies*, then the agencies can use secure summation to compute and share the sufficient statistics, following that each agency completes the analysis on its own.

3.1 Secure Regression: The Protocol

We illustrate with linear regression. Let the data consist of $p+1$ numerical attributes, so that agency j's data on its n_j subjects consist of p predictors X^j and a response y^j. Let $n = \sum n_j$ be the size of the global database. The agencies wish to fit the linear model

$$y = X\beta + \epsilon, \tag{1}$$

to the "global" data

$$X = \begin{bmatrix} X^1 \\ \vdots \\ X^K \end{bmatrix} \qquad \text{and} \qquad y = \begin{bmatrix} y^1 \\ \vdots \\ y^K \end{bmatrix}.$$

We embed the constant term of the regression in the first predictor by putting $X_1^j \equiv 1$ for all j. To illustrate the subtleties of analysis of distributed data, note that the usual strategy of centering the predictors and response at mean values does not work directly. The means in this case are the global means, which are not available, although they could be calculated with a preliminary round of secure computation.

Assume that $\mathrm{Cov}(\varepsilon) = \sigma^2 I$, in which case the least squares estimator for β is of course

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{2}$$

The global $(p+1) \times (p+1)$ matrix

$$[X\ y]^T [X\ y] = \begin{bmatrix} X^T X & X^T y \\ y^T X & y^T y \end{bmatrix}$$

[2] This vulnerability can be defeated by splitting calculation into pieces, with different orders for each, or, as in some implementations, by hiding the order from the agencies.

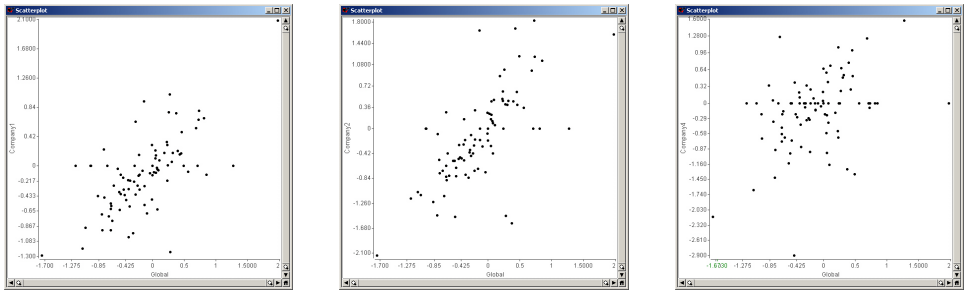

Figure 1: Scatterplots for the example discussed in §3.2. In all plots, the regression coefficients for the four-company regression appear on the x-axis. *Left:* y-axis contains regression coefficients for company 1 alone. *Center:* y-axis contains regression coefficients for company 2 alone. *Right:* y-axis contains regression coefficients for company 4 alone.

is additive over the agencies:

$$[X\,y]^T[X\,y] = \sum_{k=1}^{K}[X^k\,y^k]^T[X^k\,y^k]. \tag{3}$$

Therefore, $[X\,y]^T[X\,y]$ can be computed entrywise using secure summation, and each agency can then calculate $\hat{\beta}$ using (2).

Calculation of $\hat{\beta}$ is only part of a valid, useful regression. A variety of other objects can be calculated from $[X\,y]^T[X\,y]$, or using secure summation directly. These include the coefficient of determination R^2, the least squares estimate S^2 of the error variance σ^2, and the "hat" matrix $H = X(X^TX)^{-1}X^T$, which can be used to identify outliers (Karr et al., 2005b, 2006b). It is also possible to use the secure data integration algorithm of Karr et al. (2007), together with methods for constructing (privacy-preserving) synthetic residuals in ordinary regressions (Reiter, 2003), in order to create secure synthetic residuals (Karr et al., 2006b).

3.2 Secure Regression: Example

We illustrate with a data set of 1,318 chemical compounds (Karr et al., 2005a), in which the response is water solubility and the 91 predictors are a constant and 90 chemical features of the compounds. Four database-owning companies were created whose databases contain 499, 572, 16(!), and 231 compounds, respectively. Mimicking real-world heterogeneity, each company's database contains compounds with features that are absent from all compounds in all of the other companies' databases. This increases the incentive for companies to participate, because each can learn about the importance of features for which it has no data. Of course, company 3 has the greatest incentive to participate, since it cannot even do the regression on its own.

Figure 1 summarizes the results. The three panels are scatterplots of the 91 re-

gression coefficients for companies 1, 2 and 4 (y-axis) against the coefficients for the global (four-company) regression (x-axis). Coefficients with y-values of zero correspond to features missing from each company's database.

3.3 Other Analyses

The "additive sufficient statistics" idea is broadly applicable, and here we describe several other contexts.

Secure Contingency Tables. The algorithm for secure data integration described in Karr et al. (2007) has an important indirect application—constructing contingency tables containing counts or sums.

Let $\mathcal{D}$ be a database containing only categorical attributes $A_1, \ldots, A_J$. The associated contingency table is the J-dimensional array T defined by

$$T(a_1, \ldots, a_J) = \#\{r \in \mathcal{D} : r_1 = a_1, \ldots, r_J = a_J\}, \tag{4}$$

where each a_i is a possible value of the categorical attribute A_i and r_i is the ith attribute of record r. The J-tuple $(a_1, \ldots, a_J)$ is called the cell coordinates. The table T is a near-universal sufficient statistic, for example, for fitting log-linear models (Bishop et al., 1975).

The *sparse representation* of a table is the data structure of (cell coordinate, cell count) pairs

$$\Big\{\big(a_1, \ldots, a_J, T(a_1, \ldots, a_J)\big) : T(a_1, \ldots, a_J) \neq 0\Big\}.$$

To securely build a contingency table from databases $\mathcal{D}_1, \ldots, \mathcal{D}_K$ requires the following steps:

List of Non-Zero Cells. Use secure data integration to build the list $\mathcal{L}$ of cells with non-zero counts. The "databases" being integrated are the agencies' individual lists of cells with non-zero counts. The protocol in Karr et al. (2007) allows each agency not even to reveal in which cells it has data.

Non-Zero Cell Counts. For each cell in $\mathcal{L}$, use secure summation to determine the associated count.

Secure Maximum Likelihood Estimation. Suppose now that the agencies' databases partition a global database $\{x_i\}$ modeled as samples from an unknown density $f(\theta, \cdot)$ belonging to an exponential family:

$$\log f(\theta, x) = \sum_{\ell=1}^{\mathcal{L}} c_\ell(x) d_\ell(\theta). \tag{5}$$

Then, assuming independence, the global log-likelihood function is

$$\log L(\theta, x) = \sum_{\ell=1}^{\mathcal{L}} d_\ell(\theta) \left[\sum_{k=1}^{K} \sum_{x_i \in \mathcal{D}_k} c_\ell(x_i) \right], \tag{6}$$

where $\mathcal{D}_k$ is the database of owner k.

Assuming that the agencies have agreed in advance on the model (5), they can use secure summation to compute each of the $\mathcal{L}$ terms within the brackets in (6), and then each can maximize the likelihood function by whatever means it wishes.

3.4 Problems with the Approach

It may appear from §3.1 and §3.3 that the secure-summation-based approach is problem-free. This is anything but true.

Pre-specification. The process prevents us from being good statisticians. The analysis to be performed must be pre-specified—not only the model, but also variable transformations. There is no way to do exploratory data analysis or visualization.

No Protection Against Dishonesty. If all agencies but one are semi-honest, then that agency can not only ensure that it gets the right answer, but also that none of the other agencies get the right answer or are even aware when they don't. To see this, suppose agency j puts an incorrect value $[X^j\, y^j]^T$ in (3). Then once what the other agencies think is the correct value of $[X\, y]^T$ is calculated, it can subtract its incorrect value, add the correct value, and perform the regression. Unless the incorrect value is absurd, no other agency can detect that anything has happened.

The concept of partially trusted third parties (PTTPs) introduced in Karr et al. (2007) reduces incentives to cheat at the expense of introducing a central server performing calculations that the agencies cannot. For instance, in the setting of §3.1, a PTTP could perform the secure summation to calculate $[X\, y]^T$, but only share $\hat{\beta}$ (and related quantities derived from $[X\, y]^T$) with the agencies. Although the PTTP is trusted only with computed quantities, this may still be unacceptable to the agencies. There is no method to defeat a cheater who is content with wrecking the process and being the only one to know that it has been wrecked.

Data Heterogeneity. The notion that SMPC allows participants to learn only what is knowable from their input and the answer is most persuasive when agencies contribute approximately equally to the process. When they do not, the "information" surrendered can vary dramatically.

The simplest instance of this is when agencies have *unequal database sizes*.[3] Figure 1 provides some insight: the larger the database, the more closely the global regression resembles a company's regression.

A more pointed instance, in the example in §3.2, lies in the question "Should Com-

[3]To get a sense of this in a realistic context, populations of US states vary by a factor of 100.

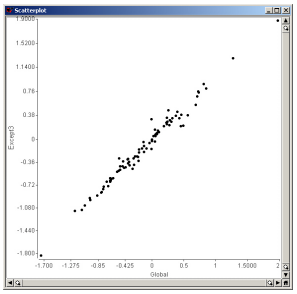

Figure 2: Scatterplot comparing global regression coefficients (x-axis) to those for the regression excluding company 3 (y-axis).

panies 1, 2 and 4 Allow 3 to Participate?" Figure 2 addresses this question by plotting the estimated coefficients from the three-company regression (1, 2 and 4) against those for the global (four-company) regression. The relationship is strong, but not perfect, which leaves the question unanswered. See **Risk-Utility Formulation** below for related discussion.

A second issue is *data heterogeneity* across agencies. It is clear that there is something fundamentally different between the top panel in Figure 3, where the three agencies possess 2-dimensional data lying in the same region, and the bottom panel, where the x-variable ranges are very different. A properly performed process would elucidate a global quadratic structure in the data that is invisible to each agency on its own, but the **Pre-specification** issue above could keep such a model from being considered.

There is a deep point here. In the "top panel of Figure 3" context, the only real benefit to the agencies is increased sample size, whereas in the "bottom panel of Figure 3" context, there is a dramatic—but given current knowledge, unattainable—benefit, if only the right analysis were done.

Problems also arise when there is differential model fit across agencies. In the regression setting, if the global fit of the global model is good, but the fit of the global model to an agency's data is poor, then it has potentially learned more about the other agencies' data than they have about its data.

The most difficult form of heterogeneity may be *differential data quality*. Data quality itself is poorly understood from a statistical perspective (Karr et al., 2006c), and to date there is no inkling about how to accommodate differences in data quality in the setting of this paper. Even the simplest problem of differential rates for missing data values is unaddressed.

Numerical and Algorithmic Issues. The protocol described in §3.1 assumes only that each agency can add.[4] By contrast, consider the problem of a secure Newton-Raphson algorithm for numerical maximization of a non-exponential family likelihood

[4]An agency unable to invert $X^T X$ may have problems, but they do not affect the other agencies.

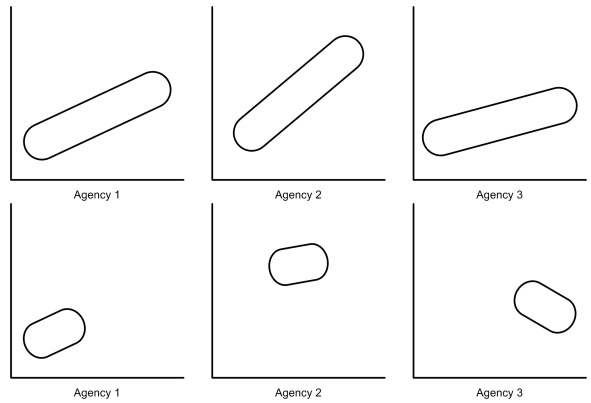

Figure 3: Homogeneous data (top) across agencies contrasted with heterogeneous data (bottom) across agencies.

function. The problem looks simple: secure summation can be used to compute the gradient vector

$$\nabla \ell(\theta_0) = \left(\sum_{k=1}^{K} \sum_{x_i \in D_k} \frac{\partial f(\theta, x_i)/\partial \theta_1}{f(\theta, x_i)}, \ldots, \sum_{k=1}^{K} \sum_{x_i \in D_k} \frac{\partial f(\theta, x_i)/\partial \theta_m}{f(\theta, x_i)} \right)_{\theta_0} \tag{7}$$

and Hessian matrix $D^2\ell(\theta_0)$ of ℓ at a given parameter value θ_0. From these, each agency can compute a Newton-Raphson step

$$\Delta\theta = -\left[D^2\ell(\theta_0)\right]^{-1} \nabla\ell(\theta_0) \tag{8}$$

and a new value $\theta' = \theta_0 + \Delta\theta$, and the process can proceed iteratively.

But there are complications. The agencies need agreed-on expressions for $\nabla\ell$ and $D^2\ell$ or shared algorithms to compute them, and these algorithms must perform identically on all machines. In addition, before the process can proceed safely from one iteration to the next, there must be a way to verify that all agencies have the same value for $\Delta\theta$ in (7). The necessary mediation mechanisms do not exist.

Non-additivity. To date, there are virtually no implemented, efficient methods for such important operations as sorting or calculating maxima. Only inefficient protocols such as bisection searches are available.

Opting Out. No methods are known that allow agencies to opt out of a secure computation *based on the results of the analysis*, should they feel that those results are too informative about their data. At some level, of course, this is a sheer impossibility: a decision that requires the results cannot be made without them. On the other hand,

it is possible to use secure summation to allow agencies to opt out beforehand on the basis of (k, p)-rules in statistical disclosure limitation (SDL) (Willenborg and de Waal, 1996, 2001) and even to make opting out anonymous. It seems likely that intermediate procedures should exist, but they remain undeveloped.

Pre-processing. The material in §3 completely omits a host of pre-processing issues related to database schemas. In particular, agencies must have the same attributes in the same units and same order, and must ensure that there are no duplicate records.

Underlying virtually all of these issues is a more fundamental shortcoming—that to date the problem formulation deviates from virtually all of modern SDL by lacking quantified measures of risk and utility.

Risk-Utility Formulation. Without a sound risk-utility formulation, our approach to secure computation is mired in a lack of clarity and inability of agencies to make decisions. There exist almost no measures (other, for example, than the relative sample size associated with a (k, p)-rule) of the risk to an agency from participating in such a protocol. Nor are there any measures—either collective or owner-specific—of the utility of an analysis. Figure 2 illustrates. Companies 1, 2 and 4 do gain *something* from allowing company 3 to participate, but how much, and does this compensate for the sizable, unquantified collective surrender of information to company 3?

4 Vertically Partitioned Data

Secure analysis of vertically partitioned data is substantially more complex than analysis of horizontally partitioned data, and less is known. To give a sense of the approaches and issues, we focus on linear regression. Explicit surrender of information is necessary, but the amount of information surrendered can be quantified and even minimized.

4.1 The Protocol

Consider agencies A, B, ..., Ω and global database $\mathbf{X}$, of size $n \times p$ partitioned vertically among them

$$\mathbf{X} = \left[\mathbf{X}^A \ \mathbf{X}^B \ \cdots \ \mathbf{X}^\Omega\right]$$

For regression, the central computational need is the $(p \times p)$-dimensional full data covariance matrix $\mathbf{X}^T\mathbf{X}$, ideally with as little surrender of information as possible.[5]

The on-diagonal blocks $(\mathbf{X}^A)^T\mathbf{X}^A$ must be computed by each agency and shared with the others. There is no alternative to this.

In Sanil et al. (2009), a secure matrix multiplication protocol is presented for computation of off-diagonal blocks $(\mathbf{X}^A)^T\mathbf{X}^B$ by pairs of agencies, who must then share the

[5]The maximum likelihood estimates of β and σ^2, as well as the standard errors of the estimated coefficients, can be obtained from the sample covariance matrix, using for example the sweep algorithm (Beaton, 1964; Schafer, 1997). The types of diagnostic measures available depend on the amount of information the owners are willing to share (Sanil et al., 2007).

result with the other agencies. Briefly, that protocol is as follows:

Step 1 Agency A generates a set of g n-dimensional vectors $\{Z_1, Z_2, \ldots, Z_g\}$ such that $Z_i^T X_j^{\mathrm{A}} = 0$ for all i and j, and sends to agency B the $(n \times g)$-dimensional matrix $\mathbf{Z} = [Z_1\ Z_2\ \cdots\ Z_g]$

Step 2 Agency B computes $\mathbf{W} = \left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^T\right)\mathbf{X}^{\mathrm{B}}$ where $\mathbf{I}$ is an $(n \times n)$-dimensional identity matrix, and sends $\mathbf{W}$ to agency A

Step 3 Agency A calculates $(\mathbf{X}^{\mathrm{A}})^T\mathbf{W} = (\mathbf{X})^T(\mathbf{I} - \mathbf{Z}\mathbf{Z}^T)\mathbf{X}^{\mathrm{B}} = (\mathbf{X}^{\mathrm{A}})^T\mathbf{X}^{\mathrm{B}}$ and shares $(\mathbf{X}^{\mathrm{A}})^T\mathbf{X}^{\mathrm{B}}$ with other agencies

How are the agencies to choose g? Consider first the two extreme cases. If $g = 0$, then $\mathbf{W} = \mathbf{X}^{\mathrm{B}}$, so agency A learns agency B's data exactly. At the other extreme, if $g = n - p$, then B knows the orthogonal complement of $\mathbf{X}^{\mathrm{A}}$ in $\mathbb{R}^n$.

To choose g, we formalize the *loss of protection* to one agency as a number of (linearly independent) constraints the other agency has on its data as a result of this process. For agency A, $\mathrm{LP(A)} = p_{\mathrm{A}}p_{\mathrm{B}} + p_{\mathrm{A}}g$, while for agency B: $\mathrm{LP(B)} = p_{\mathrm{A}}p_{\mathrm{B}} + p_{\mathrm{B}}(n - g)$. We now define the *inequity*

$$\mathrm{I}(g) = |\mathrm{LP(A)} - \mathrm{LP(B)}| = |(p_{\mathrm{A}} + p_{\mathrm{B}})g - np_{\mathrm{A}}|\,,$$

which is minimized by

$$g^* = \frac{p_{\mathrm{A}}}{p_{\mathrm{A}} + p_{\mathrm{B}}}n.$$

That is, to minimize inequity, agencies surrender information in proportion to the numbers of attributes they hold. Nothing could be more equitable.

We note in passing that a second approach to regression for vertically partitioned data (Sanil et al., 2004) requires less sharing of information, but requires that all agencies possess the response attribute y. This approach uses Powell's (derivative-free) method for quadratic optimization problems to solve directly for $\hat{\beta} = \arg\min_{\beta}\ (y - X\beta)^T(y - X\beta)$.

4.2 Problems with the Approach

Problems mirror those for horizontally partitioned data, but are generally more complex and correspondingly more poorly understood. In fact, not enough is even known to be able to identify all of the issues.

Pre-specification. Once again, the analysis to performed must be specified in advance, and must be "known" to be the right one. However, variable transformations are simple: each involves only one agency.

Analyses other than Regression. Little is known about analyses other than regression. Exceptions are Vaidya and Clifton (2002) for association rules, Vaidya and Clifton (2003) for k-means clustering and Fienberg et al. (2007) for logistic regression.

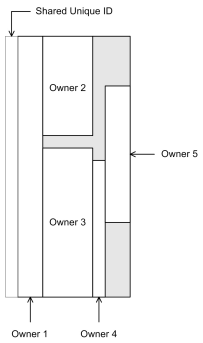

Figure 4: Example of complex data partitioning with five database owners.

Pre-processing. Records in multiple databases must be aligned with one another, which requires a common primary key.[6] The records that the databases have in common must be determined, and decisions made about how to handle incomplete records and duplicated attributes. Constraints among attributes (for instance, "gross income $\geq$ net income plus federal tax plus state tax") must be verified. Even at this stage, agencies are yielding information to one another, in the form of which records and attributes appear in whose database. Major privacy issues are involved.

Asymmetry. In §4.1, a systematic means of dealing with one form of heterogeneity is demonstrated. But there is inherent asymmetry: if only one regression is of interest, the holder of the response surrenders more information than the other agencies, especially if the coefficient of determination is high. This happens even for the secure matrix multiplication protocol, since each agency learns how well its attributes predict the other's.

Dishonesty. Multiple opportunities exist, only some of which we even know how to detect. To illustrate, if in the case of secure matrix multiplication, $(\mathbf{I} - \mathbf{Z}\mathbf{Z}^T)$ contains a column with all zeros except for a non-zero constant in one row, then agency A learns the value of agency B's data for the data subject in that row. An attribute that equals zero for all but one data subject is similarly problematic, as are records with dominant attribute values.

5 Complex Partitions

A short answer about what is known for complex data partitions—see Figure 4—that are neither purely horizontal nor purely vertical is "not much."

To begin, there may be security issues associated with knowing which agencies hold which attributes about which data subjects, and there are issues involved in even determining which subjects are common across the databases.

[6] Contrast this with §3, in which database keys are essentially immaterial.

One approach (Reiter et al., 2004; Karr et al., 2007) is to view complicated data partitions as incomplete data sets—the global database is construed as a flat file with missing values in those records not common to all parties—and then to develop secure versions of techniques used for analyzing incomplete data sets. One such technique is to specify a joint distribution for the complete data, and then to use the EM algorithm (Dempster et al., 1977) to estimate the parameters of that distribution. If the associated sufficient statistics can be calculated using SMPC, then a secure EM algorithm is feasible, as we now illustrate briefly for data following a multivariate normal distribution. For simplicity, we assume that the owners share *globally unique* identifiers of the records in their databases, in order to identify records that are common to multiple databases, and that matching on these unique identifiers can be done without error. Finally, we assume that there are no duplicate attributes.

The sufficient statistics—sums, sums of squares, and sums of cross-products of the data values—can be computed securely by the following protocol.

Let M be the number of data missingness patterns; for example, in Figure 4, $M = 5$: partitioning the attributes into four blocks (corresponding to agencies 1, {2,3}, 4 and 5), there are five patterns: blocks 3 and 4, block 3, blocks 2 and 3, block 4, and no blocks. For $m = 1, \ldots, M$, let $\mathcal{D}_m$ be the set of data elements with missingness pattern m.

To begin the secure EM protocol, the agencies group records by missingness patterns, which is possible since they have shared unique identifiers. They next compute and share two tables of sufficient statistics needed by the EM algorithm. The first table has M rows corresponding to the missingness patterns and p columns corresponding to all of the attributes in the global database. The entry for row m and column j is the sum of the observed y_j for those records with the missingness pattern associated with row m. When there are no common attributes, each sum is computed by only one owner. When there are common attributes, it is computed using secure summation.

The second table has M rows corresponding to the missingness patterns and $p(p+1)/2$ columns corresponding to the inner products of all p variables in the data set, including the sums of squares. The entry in the table for row m and the column associated with attributes (j, k) is the $\sum y_j y_k$ for those records with the missingness pattern of row m. With no common attributes, each cross-product entry in the table is derived from a single dot product involving two agencies, calculated using a secure dot product protocol (Du and Zhan, 2002; Sanil et al., 2007).

Once each agency has these two tables, it has all the information needed to run the EM algorithm (Schafer, 1997) independently of others. Further inference from the data, for example, fitting regression models, is then possible without additional error.

Beyond issues discussed already, missingness patterns associated with small numbers of attributes are problematic. In addition, the secure EM protocol does not guard against risks arising when sensitive attributes owned by different owners are nearly co-linear. A deeper difficulty is that EM algorithms are based on the assumption that the incomplete data are missing at random.

6 Conclusions and Discussion

The material in this paper epitomizes the gulf between a good idea and a sound implementation. Distributed databases are today's reality, and techniques simply must be developed to perform valid analyses on them.

Protocols based on secure summation, secure matrix multiplication, and secure dot products are comprehensible and computationally efficient. They demonstrably produce the right answer, and on the surface are safe. Working implementations, such as the NISS Secure Computation System described in Karr et al. (2007), have been constructed.

These appealing characteristics notwithstanding, current gaps in our understanding, some of which are discussed here, seem to leave us at an impasse. Official statistics agencies are, sensibly, unwilling to proceed until more is known.

This paper outlines a research agenda that will help the process move forward. As stated in §3.4, the biggest need seems to be for a risk-utility formulation that would support sound decisions by agencies. The power of such formulations in other SDL settings (Gomatam et al., 2005b,a; Karr et al., 2006a; Woo et al., 2009) is compelling reason to take on this challenge.

There remains at least one more major challenge: to link the approach with "traditional" SDL concerns about the privacy of individual data subjects. The thrust of SMPC is to protect agencies' databases from the other agencies, not to protect the data subjects. Failure to confront this problem will also be a game breaker.

Acknowledgments

The research underlying this paper was supported by NSF grants EIA–0131884 and SES–0345441 to the National Institute of Statistical Sciences (NISS) and DMS–0112069 to the Statistical and Applied Mathematical Sciences Institute (SAMSI). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

Many people have collaborated in the research reported here, most of whom have made major contributions to it. They include Stephen Fienberg, William J. Fulp, Xiaodong Lin, Yuval Nardi, Jerome Reiter, Ashish Sanil, Aleksandra Slavkovic, Francisco Vera and S. Stanley Young.

References

Beaton, A. E. (1964). The use of special matrix operations in statistical calculus. Research Bulletin RB-64-51, Educational Testing Service, Princeton, NJ.

Benaloh, J. (1987). Secret sharing homomorphisms: Keeping shares of a secret secret. In A. M. Odlyzko, ed., *CRYPTO86*, 251–260. Springer–Verlag. Lecture Notes in Computer Science No. 263.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Dempster, A., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc., Series B*, 39(1):1–38.

Du, W. and Zhan, Z. (2002). A practical approach to solve secure multi-party computation problems. In *New Security Paradigms Workshop*, 127–135, New York. ACM Press.

Fienberg, S. E., Karr, A. F., Nardi, Y., and Slavkovic, A. (2007). Secure logistic regression with distributed databases. *Bull. Internat. Statist. Inst.* Presented at the 56th Session of the International Statistical Institute, Lisbon, August 2007.

Goldwasser, S. (1997). Multi-party computations: Past and present. In *Proceedings of the 16th Annual ACM Symposium on Principles of Distributed Computing*, 1–6, New York. ACM Press.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2005a). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.*, 20(2):163–177.

Gomatam, S., Karr, A. F., and Sanil, A. P. (2005b). Data swapping as a decision problem. *J. Official Statist.*, 21(4):635–656.

Karr, A. F., Feng, J., Lin, X., Reiter, J. P., Sanil, A. P., and Young, S. S. (2005a). Secure analysis of distributed chemical databases without data integration. *J. Computer-Aided Molecular Design*, November, 2005:1–9.

Karr, A. F., Fulp, W. J., Lin, X., Reiter, J. P., Vera, F., and Young, S. S. (2007). Secure, privacy-preserving analysis of distributed databases. *Technometrics*, 49(3):335–345.

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

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

Karr, A. F., Lin, X., Reiter, J. P., and Sanil, A. P. (2005b). Secure regression on distributed databases. *J. Computational and Graphical Statist.*, 14(2):263–279.

Karr, A. F., Lin, X., Reiter, J. P., and Sanil, A. P. (2006b). Secure analysis of distributed databases. In D. Olwell, A. G. Wilson, and G. Wilson, eds., *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, 237–261. Springer–Verlag, New York.

Karr, A. F., Sanil, A. P., and Banks, D. L. (2006c). Data quality: A statistical perspective. *Statistical Methodology*, 3(2):137–173.

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

Reiter, J. P., Karr, A. F., Kohnen, C. N., Lin, X., and Sanil, A. P. (2004). Secure regression for vertically partitioned, partially overlapping data. *ASA Proc.*

Sanil, A. P., Karr, A. F., Lin, X., and Reiter, J. P. (2004). Privacy preserving regression modelling via distributed computation. In *Proc. Tenth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, 677–682. Available on-line at `www.niss.org/dgii/technicalreports.html`.

Sanil, A. P., Karr, A. F., Lin, X., and Reiter, J. P. (2007). Privacy preserving analysis of vertically partitioned data using secure matrix products. *J. Official Statist.* To appear.

Sanil, A. P., Karr, A. F., Lin, X., and Reiter, J. P. (2009). Privacy preserving analysis of vertically partitioned data using secure matrix products. *J. Official Statist.*, 25(1):125–138.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Vaidya, J. and Clifton, C. (2002). Privacy preserving association rule mining in vertically partitioned data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 639–644, New York. ACM Press.

Vaidya, J. and Clifton, C. (2003). Privacy preserving k-means clustering over vertically partitioned data. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 206–215, New York. ACM Press.

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

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

Woo, M.-J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *J. Privacy and Confidentiality*, 1(1):111–124.

Yao, A. C. (1982). Protocols for secure computations. In *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science*, 160–164, New York. ACM Press.