Using Statistics to Protect Privacy

Alan F. Karr and Jerome P. Reiter*

December 30, 2013

1 Introduction

Those who generate data—for example, official statistics agencies, survey organizations, and principal investigators, henceforth all called *agencies*—have a long history of providing access to their data to researchers, policy analysts, decision makers, and the general public. At the same time, these agencies are obligated ethically and often legally to protect the confidentiality of data subjects' identities and sensitive attributes. Simply stripping names, exact addresses, and other direct identifiers typically does not suffice to protect confidentiality. When the released data include variables that are readily available in external files, such as demographic characteristics or employment histories, ill-intentioned users—henceforth called *intruders*—may be able to link records in the released data to records in external files, thereby compromising the agency's promise of confidentiality to those who provided the data.

In response to this threat, agencies have developed an impressive variety of strategies for reducing the risks of unintended disclosures, ranging from restricting data access to altering data before release. Strategies that fall into the latter category are known as statistical disclosure limitation (SDL) techniques. Most SDL techniques have been developed for data derived from probability surveys or censuses. Even in complete form, these data would not typically be thought of as big data, with respect to *scale* (numbers of cases and attributes), *complexity* of attribute types, or *structure*: most datasets are released, if not actually structured, as flat files.

In this chapter, we explore interactions between data dissemination and big data. We suggest lessons that stewards of big data can learn from statistical agencies' experiences. Conversely, we discuss how big data and growing computing power could impact agencies' future dissemination practices. We conclude with discussion of research needed and possible visions of the future of big data dissemination.

*This work was partially supported by National Science Foundation grants CNS–1012141 and SES–1131897.

2 Experiences from Agencies

When disseminating a data set to the public, agencies generally take three steps. First, after removing direct identifiers like names and addresses, the agency evaluates the disclosure risks inherent in releasing the data "as is." Almost always the agency determines that these risks are too large, so that some form of restricted access or SDL is needed. We focus on SDL techniques here, because of the importance to researchers and others of direct access to the data. Second, the agency applies an SDL technique to the data. Third, the agency evaluates the disclosure risks and assesses the analytical quality of the candidate data release(s). In these evaluations, the agency seeks to determine whether the risks are sufficiently low, and the usefulness is adequately high, to justify releasing a particular set of altered data (Reiter, 2012). Often, these steps are iterated multiple times, for example, a series of SDL techniques is applied to the data and subsequently evaluated for risk and utility. The agency stops when it determines that the risks are acceptable and the utility is adequate (Cox et al., 2011).

To set the stage for our discussion of SDL frameworks and big data releases, we begin with a short overview of common SDL techniques, risk assessment, and utility assessment. We are not comprehensive here; additional information can be found in, for example, Federal Committee on Statistical Methodology (1994), Willenborg and de Waal (2001), National Research Council (2005, 2007), Karr et al. (2010), Reiter (2012), and Hundepool et al. (2012).

2.1 Risk Assessment for Original Data

Most agencies are concerned with two types of disclosures, namely (1) identification disclosures, which occur when an intruder correctly identifies individual records in the released data; and (2) attribute disclosures, which occur when an intruder learns the values of sensitive variables for individual records in the data (Reiter, 2012). Often agencies fold assessment of attribute risk into assessment of identification risk. For concreteness, in this chapter, we focus on data regarding individuals. In the world of official statistics, many datasets contain information on establishments such as hospitals, manufacturers and schools. Many of the problems we discuss here are significantly more challenging for establishment data (Kinney et al., 2011).

To assess identification disclosure risks, agencies make assumptions, either explicitly or implicitly, regarding what intruders know about the data subjects. Typical assumptions include whether the intruder knows that certain individuals participated in the survey, which quasi-identifying variables the intruder knows, and the amount of measurement error, or other error, in the intruder's data. For example, a common approach to risk assessment is to perform reidentification studies: the agency matches records in the original file with records from external

databases that intruders could use to attempt identifications, matching on variables common to both files such as demographics, employment histories, or education. In such studies, the information on the external files operationally defines the agency's assumptions about intruder knowledge.

Agencies are particularly concerned about data subjects that are unique in the population with respect to characteristics deemed to be available to intruders, which often are called *keys* in the SDL literature. An intruder who accurately matches the keys of a record that is unique in the population (on those keys) to an external file is guaranteed to be correct. Typically agencies only know that a record is unique on the keys in the sample. They have to estimate the probability that a data subject is unique in the population given that the subject is unique in the sample. See Skinner and Shlomo (2008) and Manrique-Vallier and Reiter (2012) for reviews of such methods. We also note that intruders who know that a particular record was in the sample can identify that record easily if it is unique in the sample on the keys.

Almost surely, the agency does not know very precisely what information intruders possess about the data subjects. Hence, and prudently, they examine risks under several scenarios, e.g., different sets of quasi-identifiers known by intruders, and whether or not intruders know who participated in the study.

2.2 Statistical Disclosure Limitation Techniques

Most public use data sets released by national statistical agencies have undergone SDL treatment by one or more of the methods below.

Aggregation. Aggregation turns atypical records—which generally are most at risk—into typical records. For example, there may be only one person with a particular combination of keys in a county, but many people with those characteristics in a state. Releasing data for this person with county indicators might pose a high disclosure risk, whereas releasing the data at the state level might not. Unfortunately, such aggregation makes analysis at finer levels difficult and often impossible, and it creates problems of ecological inferences. Another example is to report exact values only below specified thresholds, for example, reporting all ages above 90 as "90 or older." Such top coding (or bottom coding) eliminates detailed inferences about the distribution beyond the thresholds. Chopping off tails also negatively impacts estimation of whole-data quantities (Kennickell and Lane, 2006).

Suppression. Agencies can delete at-risk values, or even entire variables, from the released data (Cox, 1980). Suppression of at-risk values creates data that are not missing at random, which are difficult to analyze properly.

Data swapping. Agencies can swap data values between selected pairs of records—for example, switch counties

for two households with the same number of people—to discourage users from matching, since matches may be based on "incorrect" data (Dalenius and Reiss, 1982). Swapping at high levels destroys relationships involving both swapped and unswapped variables. Even at low levels certain analyses can be compromised (Drechsler and Reiter, 2010; Winkler, 2007).

Adding random noise. Agencies can add randomly sampled amounts to the observed numerical values, for example, adding a random deviate from a normal distribution with mean equal to zero (Fuller, 1993). This reduces the potential to match accurately on the perturbed data and changes sensitive attributes. Generally, the amount of protection increases with the variance of the noise distribution; however, adding noise with large variance distorts marginal distributions and attenuates regression coefficients (Yancey et al., 2002).

Synthetic data. Agencies can replace original data values at high risk of disclosure with values simulated from probability distributions specified to reproduce as many of the relationships in the original data as possible (Reiter and Raghunathan, 2007). Partially synthetic data comprise the original individuals with some subset of collected values replaced with simulated values. Fully synthetic data comprise entirely simulated records; the originally sampled individuals are not on the file. In both types, the agency generates and releases multiple versions of the data to enable users to account appropriately for uncertainty when making inferences. Synthetic data can provide valid inferences for analyses that are in accord with the synthesis models, but they may produce inaccurate inferences for other analyses. Despite being synthesized, synthetic data are not risk-free, especially with respect to attribute disclosure.

2.3 Disclosure Risk and Data Utility Assessment After SDL

Disclosure risk assessment. Many agencies perform re-identification experiments on SDL-protected data. In addition to matching records in the file being considered for release to external files, many agencies match the altered file against the confidential file. Agencies also specify conditional probability models that explicitly account for assumptions about what intruders might know about the data subjects and any information released about the disclosure control methods. For illustrative computations of model-based identification probabilities, see Duncan and Lambert (1986, 1989), Fienberg et al. (1997), Reiter (2005), and Shlomo and Skinner (2010).

It is worth noting that the concept of harm, such as a criminal act or loss of benefits, from a disclosure can be separated from the risk of disclosure (Skinner, 2012). SDL techniques are designed to reduce risks, not harm. Agencies may decide to take on more risk if the potential for harm is low, or less risk if the potential for harm is high. We note that agencies could be concerned about the harm that arises from *perceived* identification or at-

tribute disclosures—the intruder believes she has made an identification or learned an attribute, but is not correct—although in general agencies do not take this into account when designing SDL strategies. See Lambert (1993) and Skinner (2012) for discussion.

Data utility assessment. Data utility is usually assessed by comparing differences in results of specific analyses between the original and released data. For example, agencies look at the similarity of a set of quantities estimated with the original data and the data proposed for release, such as first and second moments, marginal distributions, and regression coefficients representative of anticipated analyses. Similarity across a wide range of analyses suggests that the released data have high utility (Karr et al., 2006). Of course, such utility measures only reveal selected features of the quality of the candidate releases; other features could be badly distorted.

The SDL literature also describes utility measures based on global functions of the data, such as differences in distributions (Woo et al., 2009). Our sense is that these methods are not widely used by agencies.

3 Current SDL and Big Data

Can typical SDL techniques be employed to protect big data? To be blunt, we believe the answer is no, except in special cases. We reach this opinion via informal, but we think plausible, assessments of the potential risk-utility tradeoffs associated with applying these methods.

3.1 Disclosure Risks in Original Files

Confidential big data carry greater disclosure risks than the typical survey sample. Often confidential big data come from administrative or privately-collected sources so that, by definition, someone other than the agency charged with sharing the data knows the identities of data subjects. This is in contrast to small-scale probability samples, which agencies believe inherently have a degree of protection from the fact that they are random subsets of the population, and since membership in a survey is rarely known to intruders. Confidential big data typically include many variables that, since the data arguably are known by others, have to be considered as keys, so that essentially everyone in the file is a population unique. Further, as the quality of administrative databases gets better with time (particularly as profit incentives strengthen their alignment with information collection), agencies cannot rely on measurement error in external files as a buffer for data protection.

3.2 Effectiveness of SDL

Because of the risks inherent in big data, SDL methods that make small changes to data are not likely to be sufficiently protective; there simply will be too much identifying information remaining on the file. This renders ineffective the usual implementations of data swapping, e.g., swapping entire records across geographies. On the other hand, massive swapping within individual variables, or even within many small sets of variables, would essentially destroy joint relationships in the data. Suppression is not a viable solution: so much would be needed to ensure adequate protection that the released data would be nearly worthless. Aggregation is likely to be problematic for similar reasons. When many variables are available to intruders, even after typical applications of aggregation, many data subjects will remain unique in the file. Very coarse aggregation/recoding is likely to be needed, which also leads to limited data utility.

One potential solution is a fully synthetic, or barely partially synthetic, data release. With appropriate models, it is theoretically possible to preserve many distributional features of the original data. However, in practice it is challenging to find good-fitting models for joint distributions of large-scale data; to our knowledge there have been only a handful of efforts to synthesize large-scale databases with complex variable types (Abowd et al., 2006; Machanavajjhala et al., 2008; Kinney et al., 2011). Nonetheless, we believe that methods for generating massive synthetic databases merit further research.

3.3 Demands on Big Data

Through both necessity and desire, analyst demands on big data will be broader than what has been dealt with to date. We know some things about utility assessed in terms of "standard statistical" analyses such as linear regressions (Karr et al., 2006), but almost nothing about utility associated with machine learning techniques such as neural networks or support vector machines, or data mining techniques such as association rules. Nor is it clear even what the right abstractions are. For instance, for surveys, SDL can to some extent be thought of as one additional source of error within a total survey error framework (Groves, 2004), allowing use of utility measures that relate to uncertainties in inferences. For big data that are the universe (e.g., of purchases at Walmart), we do not yet know even how to think about utility, let alone measure it. To illustrate, consider partitioning analyses in which the data are split recursively into classes on the basis of a response and one or more predictor variables, producing a tree whose terminal nodes represent sets of similar data points. Measuring the nearness of two trees can be challenging, making it difficult to say how much SDL has altered a partitioning analysis.

Moreover, many demands on big data will be inherently privacy-threatening. Most of today's statistical analy-

ses require only that the post-SDL data sufficiently resemble the original data in some low-dimensional, aggregate sense. For instance, if means and covariances are close, so will be the results of linear regressions. On the other hand, data mining analyses such as searching for extremely rare phenomena, like Higgs bosons or potential terrorists, require "sufficiently resemble" at the individual record level. Current SDL techniques are, virtually universally, based on giving up record-level accuracy, which reduces disclosure risk, in return for preserving aggregate accuracy, which is the current, but not necessarily the future, basis of utility.

4 Vision for The Future

We now present a vision for the future, including (i) discussions of what disclosure risk might mean and how it might be assessed with big data and big computation; (ii) how methods based on remote access and secure computation might be useful; and (iii) a vision for a big data dissemination engine involving interplay between unrestricted access, verification of results, and trusted access.

4.1 Changing Views of Privacy

It is possible, if not likely, that concomitantly with the move to big data, there also will be changes in the legal, political and social milieu within which data release lies. Of the authors of this chapter, one (AK) is a baby boomer, and the other (JR) is a genX-er. Perhaps as a result of our research on data confidentiality, our views on privacy do not differ dramatically. But, almost daily, we observe others whose views do seem quite different from ours. These include cell phone users who discuss intimate details of their lives within earshot (and "lipshot," since many of them seem aware that there are skilled lip readers everywhere) of others, social media users who seem not to realize how much privacy-compromising information a photograph can contain, and others. Whether these behaviors represent true changes in thinking about privacy, or will change as the individuals mature or societal attitudes evolve, remains to be seen. If the former, less protection of data may be required, although "who doesn't care about privacy" may be a potent form of response bias in both surveys and administrative data. If the latter, less may change.

Also unclear is the denouement of the current trend of reluctance-to-refusal by individuals to provide data to government agencies. Some of the decline in response rates is the result of privacy concerns, some is the result of everyone's increasingly complicated lives, some represents a belief that the government already has the information, and some is political opposition to *any* government data collection. It is hard not to think that response rates will continue to decline, but might they stabilize at a level at which statistical fixes still work?

What seemingly must change no matter what is the nature of the compact between data collectors and data subjects. Currently there is a major disequilibrium: official statistics agencies collect and protect much information about individuals and organizations that is readily accessible elsewhere, albeit sometimes there is a cost. Data subjects are clueless as to whether their information is protected adequately. Incentives to data subjects are seen as a means of payment for burden; subjects could (but are not now) also be compensated for (actual or risked) loss of privacy (Reiter, 2011). A supremely intriguing thought experiment is to ask "What would happen if data subjects were promised no privacy at all, and simply paid enough to get them to agree to participate?" Would data quality be destroyed? Would the cost be affordable? We do not know.

The connection between privacy and big data is likewise evolving. Answering a few questions on a survey does not generate big data, nor does it cause most people to think anew about privacy. Collecting entire electronic medical records, DNA sequences or video tapes of two years of driving (as in the Naturalistic Driving Study of the Virginia Tech Transportation Institute) may generate big data, and may change attitudes about privacy. Big data may also change what information is considered sensitive. Forty years ago, most people would have considered salary to be the most sensitive information about them. Today, a significant fraction of salaries are directly available, or accurately inferrable from, public information. Instead, medical records may be more sensitive for many people. Partly this is because (in the same way that salaries were once seen as protectable) they are perceived still to be protectable; in addition, the risk associated with knowledge of medical records may be greater (e.g., loss of insurance or a job), as well as more nebulous.

4.2 SDL of the Future: A Framework

A significant change that big data and big computing will produce is the capability to enumerate all possible versions of the original dataset that could plausibly have generated the released data. To understand what this means, we sketch here a framework for this "SDL of the future."

Let O be the original dataset and M be the released dataset after SDL is applied to O. Let $\mathcal{O}$ denote the set of all possible input data sets that could have been redacted to generate M. In general, the extent to which an analyst or intruder can specify $\mathcal{O}$ is a function of M, agency-released information about the SDL applied to O, and external knowledge. We denote this collective knowledge by (the σ-algebra) $\mathcal{K}$ and for the moment restrict it to consist only of M and agency-released information. External knowledge is addressed in §4.3.

To illustrate, suppose that O is a categorical dataset structured as a multi-way contingency table containing integer cell counts. Suppose that M is generated from O by means of suppressing low-count cells deemed to be risky, but contains correct marginal totals. In this case, additional cells must almost always be suppressed in order

to prevent reconstruction of the risky cells from the marginals. Figure 1 contains an illustration: in the table O, on the left, the four cells with counts less than 5 are suppressed because they are risky, and the cells with counts 5 and 6 are suppressed to protect them. In M, on the right, there is no distinction between the "primary" and "secondary" suppressions. Minimally, $\mathcal{K}$ consists of M and the knowledge that cell suppression was performed; $\mathcal{K}$ might or might not contain the value of the suppression threshold or information distinguishing primary from secondary suppressions. In the minimal case, $\mathcal{O}$ consists of six tables: O and the tables obtained by putting 0, 2, 3, 4, and 5 as the upper left-hand entry and solving for the other entries. We denote these by $O_0, \ldots O_5$, respectively. If the suppression threshold is known and zero is not considered risky, the first of these is ruled out because applying the rules to it does not yield M. Every one of the other four is ruled out if $\mathcal{K}$ distinguishes primary from secondary suppressions. Already one key implication for agencies is clear: the framework can distinguish what must be protected from what might be protected.

Equally important, the framework can distinguish analysts from intruders. The sardonic but apt comment that "One person's risk is another person's utility" demonstrates how subtle the issues are. Within our framework, both analysts and intruders wish to calculate the posterior distribution $P\{O = (\cdot)|\mathcal{K}\}$, but *use this conditional distribution in fundamentally different ways*.

Specifically, analysts wish to perform statistical analyses of the masked data M, as surrogates for analyses of O, and wish to understand how faithful the results of the former are to the results of the latter. (See also §4.5.) Conditional on O, the results of an analysis are a deterministic (in general, vector-valued) function $\mathbf{f}(O)$. To illustrate, for categorical data, $\mathbf{f}(O)$ may consist of the entire set of fitted values of the associated contingency table under a well-chosen log-linear model. In symbols, given $P\{O = (\cdot)|\mathcal{K}\}$, *analysts integrate* to estimate $\mathbf{f}(O)$:

$$\widehat{\mathbf{f}(O)} = \int_{\mathcal{O}} \mathbf{f}(o)\, dP\{O = o|\mathcal{K}\}. \tag{1}$$

It is important to keep in mind that $\mathcal{O}$ depends on $\mathcal{K}$, even though the notation suppresses the dependence.

To illustrate with the example in Figure 1, if $\mathcal{K}$ is only the knowledge that cell suppression was performed, then $\mathcal{O} = \{O, O_0, O_2, O_3, O_4, O_5\}$ and $P\{O = (\cdot)|\mathcal{K}\}$ is the uniform distribution on this set. By contrast, if $\mathcal{K}$ contains in addition the suppression rules, then $\mathcal{O} = \{O, O_2, O_3, O_4, O_5\}$ and $P\{O = (\cdot)|\mathcal{K}\}$ is the uniform distribution on *this* set. Finally, if $\mathcal{K}$ distinguishes primary from secondary suppressions, then $\mathcal{O} = \{O\}$.

If the analysis of interest were a χ^2 test of independence, then, in the second case, the average of the five χ^2 statistics is 34.97, and independence would be rejected. Indeed, independence is rejected for all of O, O_2, O_3, O_4 and O_5, so the analyst can be certain, even without knowing O, that independence fails.

1	18	6	25
13	5	2	20
4	1	10	15
18	24	18	60

*	18	*	25
13	*	*	20
*	*	10	15
18	24	18	60

Figure 1: *Left:* Original dataset O. *Right:* masked data set M, after cell suppression.

The point is that big computing makes this approach feasible in realistic settings.

By contrast, intruders are interested in global or local maxima in $P\{O = (\cdot)|\mathcal{K}\}$, which correspond to high posterior likelihood estimates of the original data O. In the extreme, *intruders maximize*, calculating

$$O^* = \arg\max_{o \in \mathcal{O}} P\{O = o|\mathcal{K}\}. \tag{2}$$

We do not prescribe what intruders would do using O^*, but assume only that any malicious acts would be done using O itself, for instance, re-identifying records by means of linkage to an external database containing identifiers.

This distinction allows the agency to reason in principled manners about risk and utility, especially in terms of how they relate to $\mathcal{K}$. *High utility* means that the integration in (1) can be performed or approximated relatively easily. *Low risk* means that the maximization in (2) is difficult to perform or approximate.

A central question is then: How large is the set $\mathcal{O}$ of possible values of O given $\mathcal{K}$? Of course, high utility and low risk remain competing objectives: when $\mathcal{O}$ is very large, then the maximization in (2) is hard, but so may be the integration in (1). Because of the integration in (1), it may be more natural to view $|\mathcal{O}|$ as a measure of disclosure risk than as an inverse measure of data utility.

4.3 Incorporating External Information

The framework in §4.2 meshes perfectly with a Bayesian approach to external knowledge possessed by analysts or intruders. Once $\mathcal{O}$ is known, such information exists independently of the knowledge $\mathcal{K}$ in, for instance, (1), so that it becomes completely natural to view as the product of a prior distribution on O and a likelihood function. See McClure and Reiter (2012a) for implementation of a related approach. In the example in §4.2, the prior would simply weight the elements of $\mathcal{O}$ on the basis of external knowledge.

More important from a computational perspective is that the integration in (1) can be performed by sampling from the posterior distribution $P\{O = (\cdot)|\mathcal{K}\}$, which is exactly what (Markov chain) Monte Carlo methods do!

4.4 Operational Implications

Most of today's (2013) big data are physical measurements that seem to need no SDL. There are, of course, very large transaction databases held by E-commerce websites, as well as databases containing information about telephone or E-mail communications. The extent to which any of the latter will be shared in any form is not clear. What is clear is that, in the short run at least, local storage and computing power will be supplemented or even supplanted by "cloud computing," in which, transparently to the user, data and cycles reside in multiple physical machines.

Some implications of cloud computing are troubling to official statistics agencies. They may lose control over who has physical possession of their data, over who can view the data, and how access to the data is controlled. The number of vulnerabilities increases in the cloud model, as does the possibility of secondary disclosure. In today's model, someone seeking illicitly to access Census Bureau data must penetrate Census Bureau servers, all of which are physically and electronically controlled by the Bureau. What happens if Census data might "accidentally" be seen by someone attempting to access credit card records? Can the Census Bureau legitimately promise confidentiality of records when "transparency" means lack of knowledge rather than openness? Similar, and perhaps more challenging, issues arise for licensing of datasets.

These issues notwithstanding, we expect that the data access model of the future will be to take the analysis to the data rather than the data to the analyst or the analyst to the data. There are multiple reasons for this. Truly big data are too big to take to the users. Dataset size, coupled with the current impetus for availability of research datasets, seems to demand archives that can deal with complex issues of data format, metadata, paradata, provenance and versioning. In our view, archives will also provide computational power. They will resemble today's remote access servers (Karr et al., 2010), but with vastly increased computational power and flexibility.

Construction of such archives will require addressing issues we currently choose (mostly) to bypass by limiting server capability. If the data do require protection, perhaps the most pressing challenge is query interaction: both risk and utility increase in ways we do not currently understand when multiple queries are posed to the server. Answering one query may permanently preclude answering others (Dobra et al., 2002, 2003). Many current remote access servers in effect dodge this issue by severely limiting the space of allowable queries, for instance, by forbidding high-leverage variable transformations or limiting the degree of interactions. Others include manual review of both analyses and results, a strategy that is hopelessly non-scalable. Linkage to other datasets is rarely permitted, as are exploratory tools such as visualizations. In virtually all of these instances, everything from sound abstractions to computational tools is lacking.

Because cloud data are distributed data, operational systems will require techniques for handling distributed

data. A set of techniques from computer science known generically as secure multiparty computation (SMPC) have been shown to allow analyses based on sufficient statistics that are additive across component databases (Karr et al., 2005, 2007; Karr, 2010; Karr and Lin, 2010). These analyses include creation of contingency tables, linear and logistic regression (as well as extensions such as generalized linear models) and even iterative procedures such as numerical maximum likelihood estimation using Newton-Raphson methods. For almost all other analyses, the details remain to be worked out.

4.5 Is There a Future for Microdata Releases?

In view of the discussion in §4.4, it is natural to ask whether there is a future for publicly released microdata (i.e., data on individual records). We believe that there is, but that new tools will be required to attain it.

To begin, there *is* and will remain a case for releasing microdata. Microdata are essential to the education and training of early career researchers. Historically, there has been no substitute for working directly with data, and we do not believe that this will change. (Indeed, the risk that "big data" means "disconnected from the data" is both real and disconcerting.) Perhaps more important, even skilled, mature researchers rarely know in advance which are the right questions to ask, and exploratory analyses dealing with the data themselves remain the best, if not only, path to the "right questions."

Currently available techniques for query-based analysis of distributed data using SMPC are notably poor in this respect. To illustrate, consider the example in Figure 2. There are three distributed datasets containing two variables, lying in the ranges shown. An analyst familiar with any one of the three databases would believe that the relationship between the two variables is linear, but, of course, it is quadratic instead. Existing query system models might thwart knowing the right question to ask. But, even a small sample with intensive statistical disclosure limitation (SDL) from the integrated dataset would have made the right question apparent.

The question then: if highly redacted microdata are released publicly, for example, using novel methods of generating fully synthetic data, how can an analyst know whether he or she is on the right track to the right questions, which can then be posed to an archive/server? *Verification servers* (Reiter et al., 2009; McClure and Reiter, 2012b) are one technology that offers a solution. Briefly, a verification server (VS) is a web-accessible system based on a confidential database O with an associated public microdata release M—derived from O—that

- Receives from the analyst a description of a statistical analysis Q performed on M;
- Performs the analysis on both M and O;
- Calculates one or more measures of the fidelity of $Q(M)$ to $Q(O)$;

- Returns to the analyst the values of the fidelity measure(s).

The concept is illustrated pictorially in Figure 3. When the fidelity is high, the analyst may pose the query to a server, and receive a more detailed set of results.

Verification servers also could help reduce costs of accessing servers that host confidential data. Currently, and we expect also in the future, users who want access to confidential data via virtual or physical data enclaves are vetted by the data stewards. This involves cost which often is passed to the user, for example in the form of fees to access data. With the output from a verification server, users can decide if analysis results based on the redacted data are of satisfactory quality for their particular purposes. If so, they may choose to forego the dollar and time costs of gaining access. Even users who are not satisfied with the quality of the results can benefit from starting with the redacted data. Storage and processing of big data is costly to data stewards, who likely will pass some costs to users. Analysts who have an informed analysis plan can improve their efficiency when using the server, thereby saving dollars and time.

Although attractive conceptually, verification servers remain an untested technology with both known and to-be-discovered risks. The former include risks shared with remote access servers—-unlimited and/or arbitrary queries, interaction among multiple queries, high-complexity variable transformations, subsetting of the data and intruders with extreme computational resources. Too many, or too high-precision, fidelity measures are among the latter. We do know that the latter *are* problems: if they are unaddressed, many SDL methods, including data swapping and top-coding, can be reversed (Reiter et al., 2009). At the extreme, returning to the framework in §4.2, with sufficiently many queries, sufficiently precise fidelity measures and enough computational power, O can be recovered *exactly* from M.

Archive/server-based models also seem (at least currently) to be poor at handling record linkage, except in simple cases where the linkage amounts to a database join. Knowing which variables to link with, and understanding how uncertainties are affected by linkage, require—at least in exploratory stages—actual microdata.

One item of interest in this setting is that as a means of SDL, sampling is typically seen as ineffectual, at least by itself. If the goal is to produce an analysis-capable dataset M, most records must be retained. If no other SDL is performed and this information is known, then an intruder seeking to carry out the maximization in (2) needs only to worry about the possibility that the maximizer is not in M. Typically, this would be deemed insufficient protection. On other hand, if the goal is to produce an M that allows analysts to ask the right questions, small samples, especially if accompanied by weights, may be entirely adequate.

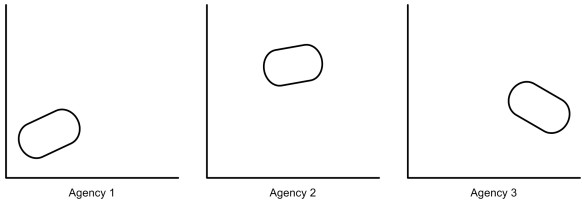

Figure 2: Three datasets, each with locally linear structure, but quadratic global structure.

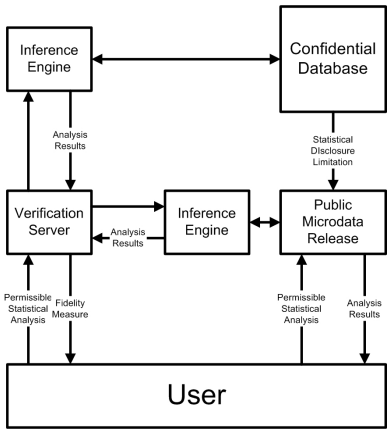

Figure 3: Pictorial schematic view of a verification server.

4.6 How Do Official Statistics Agencies Fit In?

Despite some of the challenges alluded to earlier (e.g., cloud computing), official statistics agencies are playing, and we expect will continue to play, significant roles in advancing methodology and practice for accessing big data. Many official statistics agencies that currently collect large-scale databases are experimenting with methods for providing access to these data. For example, as reported in presentations at the 2013 Joint Statistical Meetings, the Center for Medicare and Medicaid Services (CMS) has contracted with the National Opinion Research Center (NORC) to develop an unrestricted-access, synthetic public use file for Medicare claims data. This file is intended to have limited analytic utility. It exists to help researchers develop methods and code to run on the actual data. After vetting, these researchers may be approved to access the restricted data in a data enclave setting. CMS and NORC had one team develop the synthetic data sets, while another team evaluated the disclosure risks, a separation we recommend as a general dissemination practice.

The Census Bureau has forged similar partnerships with researchers in academia to develop public use products for Longitudinal Employer-Household Dynamics (LEHD) data; see `lehd.ces.census.gov` for details.

At the same time, agencies are, properly and of necessity, conservative and slow-to-change. In particular, they must deal with extremely diverse sets of data users and other stakeholders. To illustrate, agencies have been slow to

adopt multiple imputation as a means of dealing with item nonresponse, because not all users, even in the research community, are able to analyze such data. More "exotic" technologies, such as synthetic data, other Bayesian methods, and differential privacy (Dwork, 2006), will replace existing methods, if at all, only at an evolutionary pace. One promising trend, however, is increasing agency attention to the fact that most people pay heed only to the decisions based on agency data, not to the data themselves (Karr, 2012, 2013), which seems likely to yield important new insights about data utility.

5 Concluding Remarks

In spite of many steps toward wider data availability, legal, ethical, scale and intellectual property restrictions are part of the foreseeable future (Karr, 2014). "Make everything available to everyone" will not be ubiquitous, and SDL techniques are not likely to offer broadly the kind of one-off databases released by statistical agencies today. Statistical agencies already balance what to release to whom against other considerations, and this mode of thinking can, we believe, be crucial to big data.

For many big datasets, confidentiality risks of disseminating data may be so high that it is nearly impossible to share unrestricted-use microdata without massive data alterations, which call into question the usefulness of the released big data. We believe that methods for nonparametric estimation of distributions for large-scale data—a focus of significant research effort in the machine learning and statistical communities—offer potential to be converted to data synthesizers (Drechsler and Reiter, 2011). Nonetheless, unrestricted-access, big datasets probably need to take on less ambitious roles than current agency practice permits; for example, they may serve as code testbeds or permit only a limited number of (valid) analyses. Verification servers, which promise to provide automated feedback on the quality of inferences from redacted data, could enhance the usefulness of such datasets, allowing users to determine when they can trust results and when they need to accept the costs of applying for access to the confidential data. Highly redacted datasets also should help users of remote query systems to identify sensible queries.

To conclude, we believe a way forward for big data access is an integrated system including (i) unrestricted access to highly redacted data, most likely some version of synthetic data, followed with (ii) means for approved researchers to access the confidential data via remote access solutions, glued together by (iii) verification servers that allow users to assess the quality of their inferences with the redacted data so as to be more efficient with their use (if necessary) of the remote data access. We look forward to seeing how this vision develops.

References

Abowd, J., Stinson, M., and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at `http://www.census.gov/sipp/synth_data.html`.

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

Cox, L. H., Karr, A. F., and Kinney, S. K. (2011). Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act (with discussion). *Int. Statist. Rev.*, 79(2):160–199.

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

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

Dobra, A., Karr, A. F., and Sanil, A. P. (2003). Preserving confidentiality of high-dimensional tabular data: Statistical and computational issues. *Statist. and Computing*, 13(4):363–370.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105:1347–1357.

Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*, 55:3232–3243.

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81:10–28.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7:207–217.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I., editors, *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer–Verlag, Berlin.

Federal Committee on Statistical Methodology (1994). *Report on Statistical Disclosure Limitation Methodology*. US Office of Management and Budget, Washington. Statistical Policy Working Paper 22.

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

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

Groves, R. M. (2004). *Survey Errors and Survey Costs*. Wiley, New York.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte-Nordholt, E., Spicer, K., and de Wolf, P.-P. (2012). *Statistical Disclosure Control*. New York: Wiley.

Karr, A. F. (2010). Secure statistical analysis of distributed databases, emphasizing what we don't know. *J. Privacy and Confidentiality*, 1(2):197–211.

Karr, A. F. (2012). Discussion on statistical use of administrative data: Old and new challenges. *Statist. Neerlandica*, 66(1):80–84.

Karr, A. F. (2013). Discussion of five papers on "Systems and Architectures for High-Quality Statistics Production". *J. Official Statist.*, 29(1):157–163.

Karr, A. F. (2014). Why data availability is such a hard problem. *Statistical Journal of the International Association for Official Statistics*. To appear.

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., Kinney, S. K., and Gonzalez, Jr., J. F. (2010). Data confidentiality—the next five years: Summary and guide to papers. *J. Privacy and Confidentiality*, 1(2):125–134.

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

Karr, A. F. and Lin, X. (2010). Privacy-preserving maximum likelihood estimation for distributed data. *J. Privacy and Confidentiality*, 1(2):213–222.

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

Kennickell, A. and Lane, J. (2006). Measuring the impact of data protection techniques on data utility: Evidence from the Survey of Consumer Finances. In Domingo-Ferrar, J., editor, *Privacy in Statistical Databases 2006 (Lecture Notes in Computer Science)*, pages 291–303. New York: Springer-Verlag.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review*, 79:363–384.

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

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering*, pages 277–286.

Manrique-Vallier, D. and Reiter, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*, 107:1385–1394.

McClure, D. and Reiter, J. P. (2012a). Differential privacy and statistical disclosure risk measures: An illustration with binary synthetic data. *Transactions on Data Privacy*, 5:535–552.

McClure, D. and Reiter, J. P. (2012b). Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality*, 4:1:Article 8.

National Research Council (2005). *Expanding Access to Research Data: Reconciling Risks and Opportunities*. Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington DC: The National Academies Press.

National Research Council (2007). *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data, Committee on the Human Dimensions of Global Change, Division of Behavioral and Social Sciences and Education. Washington DC: The National Academies Press.

Reiter, J. P. (2005). Estimating identification risks in microdata. *Journal of the American Statistical Association*, 100:1103–1113.

Reiter, J. P. (2011). Commentary on article by Gates. *Journal of Privacy and Confidentiality*, 3:Article 8.

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly*, 76:163–181.

Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis*, 53:1475–1482.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102:1462–1471.

Shlomo, N. and Skinner, C. J. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Annals of Applied Statistics*, 4:1291–1310.

Skinner, C. (2012). Statistical disclosure risk: Separating potential and harm. *International Statistical Review*, 80:349 – 368.

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

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Winkler, W. E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Technical report, U.S. Census Bureau Research Report Series, No. 2007-21.

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

Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In Domingo-Ferrer, J., editor, *Inference Control in Statistical Databases*, pages 135–152. Berlin: Springer-Verlag.