Does Big Data Change the Privacy Landscape? A Review of the Issues

Sallie Ann Keller, Stephanie Shipp, and Aaron Schroeder

Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech, Arlington, Virginia 22203; email: sallie41@vbi.vt.edu

Annu. Rev. Stat. Appl. 2016. 3:12.1–12.20

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-041715-033453

Keywords

confidentiality, statistical disclosure limitation, identification, record linkage, trust but verify

Abstract

The current data revolution is changing the conduct of social science research as increasing amounts of digital and administrative data become accessible for use. This new data landscape has created significant tension around data privacy and confidentiality. The risk–utility theory and models underpinning statistical disclosure limitation may be too restrictive for providing data confidentially owing to the growing volumes and varieties of data and the evolving privacy policies. Science and society need to move to a trust-based approach from which both researchers and participants benefit. This review discusses the explosive growth of the new data sources and the parallel evolution of privacy policy and governance, with a focus on access to data for research. We provide a history of privacy policy, statistical disclosure limitation research, and record linkage in the context of this brave new world of data.

THE ALL DATA REVOLUTION

Until recently, national or federal statistical systems have been the primary source of data to understand the social condition in the United States. This data landscape has been changing for some time, and many feel we are in the middle of a data revolution wherein state and local government administrative data, as well as private sector data such as Twitter messages, have become widely available to analysts (Keller et al. 2012, NRC 2013). This shift has consequences for the privacy backdrop of data access. This review discusses the explosive growth of these new data sources and the parallel evolution of privacy policy and governance, with a focus on access to data for research. Further, it provides a history of privacy policy and statistical disclosure limitation research in the context of an evolving data landscape. The intent is to provide enough background to help the reader advance the privacy conversation in the context of new approaches based on statistical theory and methods that are needed in this brave new world of data.

The National Research Council's *Private Lives and Public Policies* (NRC 1993) played an important role in contributing to the US federal statistical system's approach to privacy and the collection and dissemination of statistical and research data. The recommendations in that report proposed changes in privacy policies, many of which were implemented in subsequent years. Today, the "all data" revolution (Lazer et al. 2014) that was first identified as the "big data" revolution (Manyika et al. 2011) has changed the privacy discussion as the availability of digital and administrative data has made massive flows of data available for research and other uses.

This all data revolution is rapidly generating petabytes of data, with estimates that nondesigned data collection now accounts for 70% of all data; this percentage is expected to grow rapidly every year to eventually account for virtually all data (FCC 2011). As Ken Prewitt, Carnegie Professor of Public Affairs at Columbia University and former director of the US Census Bureau, noted at a recent meeting at the National Academy of Sciences, "By the time you finish reading this sentence, there will have been 219,000 new Facebook posts, 22,800 new tweets, 7,000 apps downloaded, and about $9,000 worth of items sold on Amazon" (VASEM 2014).

The all data revolution is changing the focus of the privacy discussion from the masking and suppression of data in order to maintain confidentiality, to trust, policy, and governance—indeed, itself a revolution in thinking about privacy. Privacy and confidentiality are defined as follows (NRC 2007b):

- Privacy refers to the amount of personal information individuals allow others to access about themselves.
- Confidentiality is the process that data producers and researchers follow to keep individuals' data private.

The availability of massive amounts of data and the ability to find associations across many data sources make it impossible to achieve assurances of complete confidentiality if the data are to still be useful for research. Changes in cultural norms and governance frameworks for data access are emerging that rely on a trust-but-verify approach, balancing trade-offs of privacy with societal gains from the use of the data (Erlich et al. 2014, NRC 2013, PCAST 2014). Yet this revolution is at odds with existing legislation stating that any potential disclosure risk is unacceptable.

Sources of Data

Traditionally, the data used to study society are from statistically designed data collections, usually surveys and experiments. To date, statistical disclosure limitation theory and methodology have been developed almost exclusively for such data collections. The typical approach is to control access and use by masking the data through one or more methods, such as cell suppression, random error, or bottom- or top-coding (e.g., setting all income over a certain limit to that limit or substituting the mean of all top-coded values). Researchers and data producers, especially those who collect and disseminate federal statistical data, have always worried about controlling both access and use of the data. However, until recently, most data confidentiality methods and systems have controlled use by controlling access.

These traditional data are now being augmented with two distinct new data sources, digital data, such as global positioning systems (GPS), embedded sensors, and social media exchanges, and administrative data, such as government, commercial, and financial transactions. Not only are these new sources not statistically designed data collections, but they are not designed data collections of any sort. They come in many forms—numbers, text, images, and sound—and can be transmitted autonomously and captured through program applications. All of this provides large volumes of data in near real time.

Digital data are captured on an ongoing basis through information communication technologies. These data come from all varieties of sensors, from mobile and wearable devices, or simply via the Internet, such as through searches and web crawling and scraping, generating volumes of data collected opportunistically on topics of one's choosing. Research using data from crowdsourcing organizations, such as Amazon Mechanical Turks (Mason & Suri 2012) or Jana.com (Berinsky et al. 2012, Bohannon 2011), exists because such organizations offer cheap venues to collect data. The President's Council of Advisors on Science and Technology refers to these data as information that is "born digital" (PCAST 2014).

Administrative data are data collected for the administration of an organization or a program. These data can be leveraged to augment statistically designed data collections in the development of estimates and trends (Burwell 2014). Examples of these include Internal Revenue Service (IRS) data for individuals and businesses; Social Security earnings records; Medicare and Medicaid health utilization data; 911, emergency management services, and property tax data from local governments; credit card data; salary data for state employees; and taxi data (Bohannon 2015, Goroff 2015). It is only incidental to this main purpose that such data may provide useful and detailed information that could complement or supplement survey data. Although these data may be coming from business enterprise systems or official records of insurance claims or purchases, they have unknown statistical properties such as bias, sources of error, and lack of representativeness. In some cases, these statistical properties may be knowable but simply have not been well studied (NRC 2013).

Advantages and Disadvantages of Nonstatistically Designed Sources of Data

These nonstatistically designed sources of data are intoxicating in that they provide easily accessible and often inexpensive information about individuals, businesses, and society. They offer possibilities for studying behavior and social drivers of population attributes and characteristics at a finer level of geographic and demographic resolution and in more frequent time intervals than do survey and census data (Braaksma & Zeelenberg 2015). They hold the promise of understanding human interactions at a societal scale, within a context of rich spatial and temporal dynamics, and for detecting complex interactions and nonlinearities among variables (Agafiţei et al. 2015). In contrast to designed data collection, the unit cost is inexpensive. For these data sources, the clear cost drivers are no longer the data collection methods themselves but rather the development and execution of the data analytics, a significant game changer in how to think about social and decision informatics.

Conclusions derived from designed experiments and statistical surveys are extremely well understood. Decades of research have gone into characterizing and understanding all sources of error and bias associated with the results, even if they are not always fully measured. In traditional studies, how the data are produced is every bit as important as the findings derived from the data. Conducting statistically designed studies, such as randomized clinical trials or large surveys, has become increasingly expensive and time-consuming, and the latter face increasing nonresponse and nonparticipation. This leads to wide interest in the use of other data sources to augment or complement (or even perhaps replace) traditional studies. However, each of these new data sources has opportunities and challenges.

Use of administrative data is not as well understood as the use of survey data. The opendata.gov movement has made some of these data available for research, such as data on healthcare utilization by Medicare and Medicaid patients from the Centers for Medicare and Medicaid Services (Brennan et al. 2014). These data can seem massive in size for a researcher used to traditional sources of data.

Administrative data is also increasingly in use and available in the private sector. Data aggregators combine data from multiple sources; for example, Truven Health amasses health insurance claims and CoreLogic aggregates housing and property data across multiple geographic areas. The data come with little to no documentation about coverage, representativeness, bias, and gaps in the data. If these data are adapted for research, they may present time comparability problems, because over time companies (or reporting units) may merge or change focus, leading to longitudinal reporting gaps.

Even less is known about the statistical properties of digital data. Reverse engineering is one approach used to assess their representativeness and accuracy related to the research questions of interest. On the other hand, these data may provide insights into social behaviors that may not be possible to capture with time-constrained statistical surveys, designed experiments, or administrative data. A clear understanding of the bias trade-offs is needed.

Examining the advantages of these new sources of data also unveils their disadvantages. The data are more granular, but one has little control over the quality of the data. There is also a lack of transparency about the quality of the data, especially when it is privately owned, such as Google searches or data from credit card transactions. Importantly, from a statistical perspective, there is a lack of information about the representativeness and error structure of the data (NRC 2013).

Privacy concerns can arise from the massive amounts of collected data, the changes in and context of how these data will be used, and the potential for overcollection (PCAST 2014). Overcollection of digital data occurs when data are collected that are not related to the stated research purpose. Examples of overcollection may include collecting mouse clicks, taps, swipes, or keystrokes; email messages that contain information about the sending and destination addresses; GPS location data; metadata associated with phone calls (e.g., the numbers dialed from or to, the time and duration of calls); data associated with most commercial transactions [e.g., credit card swipes, barcode reads, reads of radio-frequency identification (RFID) tags used for antitheft and inventory control]; and data from cars, televisions, appliances, and the like, also known as the Internet of Things.

When using these new sources of data, it is useful to step back and reflect on what has traditionally been done in social and behavioral science research. With the exception of some researchers designing their own data collection instruments (most particularly social experiments), the conduct of research has been guided by instruments, such as official surveys and censuses, to gather data every decade, quarter, year, or month. Self-reported small samples of interview data are used in daily, weekly, or monthly time frames. Experiments have been used to provide more in situ behavioral data, but the experimental situations may not represent reality as they are often contrived and the sample size is still small.

The new modalities of data flows allow social and behavioral data to be captured every second or more frequently for years, while the individuals are in the act of behaving. These data may give social scientists the ability to observe behavior at a new level of resolution. This is analogous to acquiring a new scientific instrument, a confocal microscope for the social condition (Keller & Shipp 2016), with an increased level of sensitivity over previous instruments. The important point is that this new instrument does not make the previous instruments (e.g., statistically designed data collection) obsolete, but rather allows scientists to observe events that could not previously be resolved at a greater level of detail or granularity.

EVOLUTION OF PRIVACY POLICY

Researchers and data producers, especially those who collect and disseminate federal statistical data, have always worried about controlling collection of, access to, and use of the data. The level of detail that a person requires for privacy varies by the individual. In a research context, for nongovernment data collection, the Institutional Review Board (IRB) process requires that people participate on a voluntary basis, provide only the information that they want to, and are able to end their participation at any time (NRC 2007b). The IRB role is to review the survey (or study process) to ensure the rights and safety of human subjects participating in research. IRB review is required by law for federally funded studies.

The processes that data producers use to protect the data define data confidentiality. These processes include an "ethical responsibility to avoid intrusion on participants' privacy and to minimize the likelihood of harm from the disclosure of private information" (NRC 2007b, p. 129). More broadly, confidentiality procedures are designed to protect data on individual persons, households, companies, and other institutions (NRC 2007a). Researcher access to confidential data is frequently provided through (*a*) the release of microdata files in which some data are suppressed, top-coded, swapped, or perturbed; (*b*) query-based remote access; or (*c*) on-site access such as the Census Bureau Research Data Centers (RDCs).

Historically, the US federal statistical system focused on protection and security, accountability, and rights and responsibilities for using personal data. These privacy principles enunciated by the Organisation for Economic Co-operation and Development(OECD) were developed in the 1970s, published in 1980 (OECD 1980), and updated in 2013 (OECD 2013). In the United States, similar principles were indirectly adopted through a potpourri of laws that require statistical agencies to protect the identity (privacy) of respondents. **Table 1** presents some of the key laws governing statistical agencies.

The Privacy Act of 1974 limits the collection of unnecessary information and the disclosure of identifiable information to third parties. When these laws and regulations are applied, new challenges are created. For example, because of dozens of exemptions, the Privacy Act fails to distinguish statistical and research purposes from administrative purposes in restricting access to individually identifiable records. This poses a major obstacle for data users who want to analyze such information for statistical or research purposes (NRC 1993).

The Freedom of Information Act (FOIA) of 1966 allows the public to request access to records maintained by federal agencies unless the request for access falls within one of nine specific exemptions (US DOJ 2015). FOIA presents challenges similar to those of the Privacy Act in that the exemptions may lead to the release of confidential data not covered by FOIA (NRC 1993). Statistical data cannot be accessed and are exempt based on legislation that protects the confidentiality of respondents (US DOJ 2015). However, statistical records maintained by federal agencies, even those developed by private parties, are subject to disclosure under the Act if the release of the data are not otherwise exempt.

Title 13 of the United States Code provides the authority for the Census Bureau to collect data from and publish statistics about individuals, households, and businesses (US Census Bur. 2015b).

Table 1 Selected US federal statistical system privacy laws[a]

Legal acts and titles related to privacy and confidentiality	Description
Privacy Act of 1974 (Public Law 93-579)	Requires that federal agencies (*a*) grant individuals access to their identifiable records maintained by the agency, (*b*) ensure that existing information is accurate and timely and limit the collection of unnecessary information, and (*c*) limit the disclosure of identifiable information to third parties.
Freedom of Information Act of 1966 (Public Law 89-487)	Allows public access to records maintained by federal agencies unless the request for access meets one of nine specific exemptions. Statistical records maintained by federal agencies, even those developed by private parties, are subject to disclosure under the act if not otherwise exempt.
Title 13, US Code (codified 1954; amended 1962, 1990)	Defines the authority under which the US Census Bureau collects information to produce statistics about individuals, households, and businesses. ■ Bureau employees are sworn to protect confidentiality of these data. ■ Private information is never published. ■ Violating the law is a serious federal crime.
Title 26, US Code: Internal Revenue Code (codified 1939, amended 1954, 1986)	Applies to the statistical work conducted by the US Census Bureau's collection of Internal Revenue Service data about households and businesses. "Publication of all statistical products by the Census Bureau . . . are subject to disclosure avoidance procedures . . . The Census Bureau's main computer system that stores and processes the Personally Identifiable Information (PII) resides behind the Census Bureau firewall(s). . . . All activity on the system is recorded in security audit logs that are reviewed on a regular basis by designated personnel" (US Census Bur. 2015c).
Confidential Information Protection and Statistical Efficiency Act of 2002 (Public Law 107-347)	"Subtitle A protects information that is acquired for exclusively statistical purposes under a pledge of confidentiality . . . [and] applies to all Federal agencies that acquire information under these carefully prescribed conditions. The protection of information collected under this law is supported by a penalty of a Class E Felony for a knowing and willful disclosure of confidential information. This includes imprisonment for up to five years and fines up to $250,000" (OMB 2007, p. 4).

[a]Presented in the order discussed in the text.

Title 13 requires protection of the collected information: The Census Bureau cannot and does not publish private information, and this information cannot be used against any respondent by a government agency or court. Bureau employees are sworn to protect the confidentiality of the data and there are steep penalties for violating this law.

Title 26 applies to IRS employees, other federal agencies, and researchers with access to tax data. Title 26 allows the IRS to provide federal tax returns and tax return information to the Census Bureau and other agencies for statistical purposes (US Census Bur. 2015c).

Nine years after the National Research Council's *Private Lives and Public Policies* (NRC 1993) recommended new legislation, Congress enacted legislation to create the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002 to provide a unified approach to confidentiality at federal statistical agencies. CIPSEA strengthened the protections afforded to confidential statistical information for agencies such as the Bureau of Labor Statistics and the Bureau of Transportation Statistics but kept in place the stringent laws governing many other agencies (Wallman & Harris-Kojetin 2004). For example, four separate laws cover the protection of the confidentiality of individually identifiable information collected by the National Center for Education Statistics (NCES): the Privacy Act of 1974, the Education Sciences Reform Act of 2002, the USA Patriot Act of 2001, and the E-Government Act of 2002. The Patriot Act reduced the confidentiality protection for education records for investigations and prosecution of specific crimes or acts of terrorism (US DOE 2015).

Similarly, the legislation that created the National Center for Health Statistics (NCHS) prohibits the NCHS from using any personal information for any purpose other than what was described to survey participants and from sharing that information with anyone not clearly mentioned to them. In addition, the NCHS must follow the Privacy Act and CIPSEA legislation (CDC 2015a). In a similar way, the Census Bureau's Privacy Principles combine and apply the provisions of Titles 13 and 26 across all surveys, requiring the protection of confidential information, implementation of statistical safeguards, and data stewardship (US Census Bur. 2015a).

Other laws, such as the Paperwork Reduction Act of 1980 (updated in 1995), Family Educational Rights and Privacy Act of 1974, and Health Insurance Portability and Accountability Act (HIPAA) of 1996, also restrict the use and release of data (IOM 2009), which is important for protecting the data but also very restrictive when using the data in research.

EVOLUTION OF STATISTICAL DISCLOSURE LIMITATION RESEARCH

In the 1960s, the Census Bureau released some of the first microdata files for research purposes (US Census Bur. 2015d). That release carried with it the awareness that researchers would want more and more data, leading to increased complexity in keeping the data confidential. By the late 1970s, researchers created methods to deliberately safeguard data confidentiality through the use of data swapping (Dalenius & Reiss 1982) and cell suppression (Cox 1982). By the mid-1980s, a landmark publication emerged that would codify the field of statistical disclosure limitation, providing a theoretical foundation for assessing disclosure control policies (Duncan & Lambert 1989). The general idea is that the amount of information a data user has before and after the release of data is quantified through prior and predictive distributions. This allows for disclosure policies to be mathematically characterized by placing data release constraints on the predictive distribution.

Duncan & Lambert (1986) extended the disclosure-limiting framework to the release of microdata. At the time, owing to the formalization of statistical disclosure limitation and the growth in network and database technologies, some researchers were beginning to introduce the notion of remote access to federated confidential data (Duncan & Lambert 1989, Keller-McNulty & Unger 1998). The early 1990s brought some harmonization between the computer science inferential security field and statistical disclosure limitation (Keller-McNulty & Unger 1993, NRC 1993).

This innovation opened opportunities to access new and more data sources and simultaneously created tension between researchers and data providers, for example, the federal statistical system. Tensions between new sources of data and privacy have occurred throughout American history. These tensions are primarily addressed through the adoption of principles and legislation but often lag in being updated or changed as new sources of data emerge. Many proposed institutional solutions involve establishing tiers of risk and access by developing data-sharing protocols that match the level of access to the risks and benefits of the planned research (NRC 2007b). As described below, the statistical sciences community mathematically formalized this concept.

The late 1990s marked the development in the statistical disclosure limitation literature of the concept of risk–utility trade-off for defining data confidentiality polices (Duncan & Fienberg 1997). The framework balances access to data and utility of the analytics and hence is an integrated model for data access control and use. This theory is refined by Duncan et al. (2001), applied to statistical databases (Duncan et al. 2004), embedded in remote access servers (Gomatam et al. 2005), placed in an information game-theoretic framework (Keller-McNulty et al. 2005), and applied broadly to disclosure-limiting techniques of recoding, top-coding, and swapping (Reiter 2005a).

In the 1980s and 1990s, remote access systems allowing researchers controlled access to confidential statistical data began to be created. The history of remote access systems started with the Luxembourg Income Study, which began in 1983, and was extended to the Luxembourg Employment Study in 1994 (Schouten & Cigrang 2003). The essence of these early systems lay in queries sent via email or over the web resulting in agencies preparing responses, typically aggregate tabulations, ensuring confidentiality of the underlying data. For example, the Integrated Public Use Microdata Series (IPUMS)-International project set up in 1999 allowed access to the census data of many different countries. This was done through web requests that produced properly protected tabulations. In 1998, progress on sequential query access to confidential data gained traction with the emergence of the first prototype (Keller-McNulty & Unger 1998). In the late 1990s to early 2000s, the National Institute of Statistical Science developed a system under the Digital Government Strategy. This system monitored the sequence of queries made and applied increased confidentiality controls as warranted to project the intersection of the queries (Schouten & Cigrang 2003). Another application of this approach is proposed in combination with the use of synthetic microdata and access to gold standard data through a virtual RDC (Abowd et al. 2004). The problem with these approaches is that they simply do not scale, and currently there may not always be good technical approaches to weigh the risk–utility trade-off for big data applications.

In the mid-1990s, the Census Bureau created the first RDC in Boston (US Census Bur. 2013). RDCs follow the traditional approach to controlling use through access. The microdata are not specifically controlled, such as by cell suppression or masking, so researchers must conduct their research at a restricted facility, and their output must be reviewed and deemed not to violate confidentiality. The Census RDCs control access by controlling who has access and requiring researchers to be subject to the same rules and penalties as Bureau employees. The Census Bureau uses IRS data for conducting censuses and related statistical activities authorized by law. The Bureau has set up 20 RDCs (with partnerships at 50 locations) to allow researcher access to microdata in a secure way; however, as noted elsewhere in the review, this approach is restrictive.

In 2006, the National Opinion Research Center (NORC) at the University of Chicago launched a data enclave to provide remote access to confidential microdata on businesses for the US federal statistical agencies (Campbell et al. 2009). It uses a portfolio approach to data access that includes some statistical protection (mainly deleting obvious identifiers), screening of researchers, training for researchers in legal and ethical confidentiality requirements, and secure onsite and remote access. One of the key features of the enclave is a collaborative environment within which researchers can share knowledge about the data and build on each other's work (Lane & Shipp 2008). The NORC data enclave is one of the first mechanisms that allows for simultaneous control of access and use and is an early example of a privacy-enabling interface design. Since then, statistical agencies have implemented remote access RDCs, such as the one sponsored by the NCHS, to access their data and those of other sources across the Department of Health and Human Services (CDC 2015b). These examples of remote access are important steps in the right direction, but they still work in very restricted contexts.

During this period of providing access to microdata through secure environments, researchers began exploring synthetic data alternatives. A step beyond masking (altering) the original data is to generate synthetic data as a surrogate for the original data (Abowd & Lane 2003, Abowd et al. 2009, Raghunathan et al. 2003, Reiter 2005b). This work borrows from multiple imputation research, originally developed to address missing data (Rubin 1987, 1996). The early applications of imputation for disclosure limitation ranged from imputing a subset of sensitive variables in the data (partially synthetic data) to using multiple imputations to fully synthesize multiple realizations of the data set. These imputations are done to preserve important features and structure in the data. Analysis methods that can be appropriately applied to the synthetic data are still limited (Reiter 2009). However, progress is being made both in characterizing the disclosure risks associated with synthetic data (Reiter et al. 2014) and in more complex applications (Schneider & Abowd 2015).

Even with the technical evolution and options for providing confidential access to data for research, current privacy laws still focus on restricting data access, are confusing, often have unintended consequences, and are based on out-of-date ideas about anonymization and identifiability (Goroff 2015). An Institute of Medicine study concludes that the HIPAA Privacy Rule does not protect privacy as well as it should, and that as currently implemented, it impedes important health research (IOM 2009). These approaches focus on information access, defined as how information comes to be known (Kagal & Abelson 2010). In addition, the current notice and consent framework is "a central pillar of how privacy practices have been organized for more than four decades" (PCAST 2014). The all data revolution changes this premise: "The notice and consent is defeated by exactly the positive benefits that big data enables: new, nonobvious, unexpectedly powerful uses of data" (PCAST 2014).

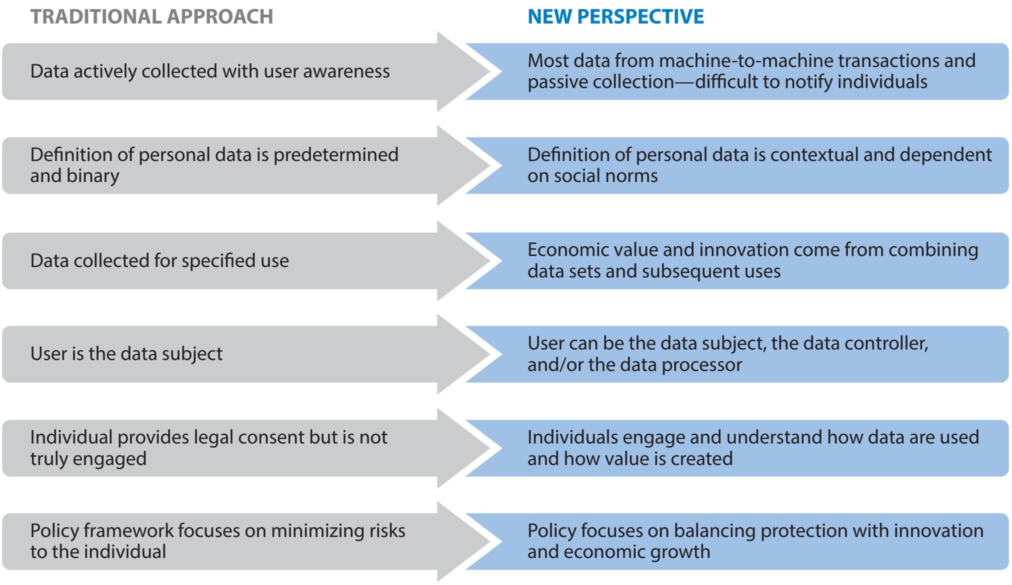
Past conflicts between privacy and new technology have generally related to what is now termed small data, the collection and use of data sets by private- and public-sector organizations for which the data are disseminated in their original form or analyzed by conventional statistical methods (PCAST 2014). Today's concerns about big data reflect the substantial increases in the amount of data being collected and associated changes, both actual and potential, in how they are used (PCAST 2014). Government regulation can provide cues that increase trust and reduce consumer concern about privacy and result in sensible behavior (Acquisti et al. 2015).

The Consumer Bill of Rights (CBR), updated in 2012, begins to set the stage for the proposed trust-based approach described below. The principles for the CBR require companies to respect data and only use it in the context in which the data are collected, to focus collection to meet needs, to provide secure handling of data, and to be accountable for ensuring they meet these requirements. The CBR also gives consumers the right to exercise control over what personal data companies collect from them and how they use the data, transparency about privacy and security practices, and the right to access and to also correct the data in a usable format. However, the CBR is restrictive in that it limits data collection to immediate uses (EOP 2014, Landau 2015).

In the end, current approaches that give the appearance of promoting accuracy and privacy are disproved by new research that demonstrates that anonymization and current suppression methods often do not work (de Montjoye et al. 2014, Goroff 2015, PCAST 2014). Many argue that the traditional methods of protecting privacy are obsolete and run counter to protecting civil liberties (Landau 2015, Mundie 2014). Specific to statistical disclosure-limiting techniques, Karr & Reiter (2014) clearly articulate that the small changes traditionally made through statistical controls are insufficient to protect data in the context of big data. They go on to say that applying current methodology will require massive swapping, suppressions, or perturbations, rendering the data useless for statistical purposes and research. Many seem to agree that new approaches with corresponding regulation are needed.

PRAGMATIC APPROACHES TO PRIVACY: A TRUST-CENTRIC PRIVACY MODEL

> Indeed, there is hardly any part of one's life that does not emit some sort of "data exhaust" as a byproduct. And it has become virtually impossible for someone to know exactly how much of his data is out there or where it is stored.... The time has come for a new approach: shifting the focus from limiting the



TRADITIONAL APPROACH	NEW PERSPECTIVE
Data actively collected with user awareness	Most data from machine-to-machine transactions and passive collection—difficult to notify individuals
Definition of personal data is predetermined and binary	Definition of personal data is contextual and dependent on social norms
Data collected for specified use	Economic value and innovation come from combining data sets and subsequent uses
User is the data subject	User can be the data subject, the data controller, and/or the data processor
Individual provides legal consent but is not truly engaged	Individuals engage and understand how data are used and how value is created
Policy framework focuses on minimizing risks to the individual	Policy focuses on balancing protection with innovation and economic growth

Figure 1

The World Economic Forum has provided a mapping between the traditional approach to data collection and protecting anonymity and a new perspective that aligns to the all data revolution. The traditional approach protects the data elements in an effort to minimize individual disclosure risk. The new perspective encourages policy that focuses on the use and economic value of the data. Figure adapted with permission from Kalapesi (2013).

> collection and retention of data to controlling data at the most important point – the moment when it is used. (Mundie 2014)

As the quote from Mundie notes, new approaches to privacy are being proposed that take into account the all data revolution. **Figure 1** provides a summary of the traditional approach (described above) and the new perspective, discussed by the World Economic Forum (Kagal & Abelson 2010, Kalapesi 2013, Schwab et al. 2011). Since then, other reports and articles have further developed these ideas, providing approaches for implementation (Erlich et al. 2014, Landau 2015, PCAST 2014). The fundamental premise is that new, less prescriptive approaches to privacy are needed to maximize the benefits of using big data while minimizing harm.

This new perspective is that the use of multiple sources of personal data requires a balance between protecting individuals' privacy and using the data in research to achieve economic growth and innovation. Characteristics of this new approach are flexibility and adaptability. This will require a shift from governing the collection and management of data to understanding the context (Nissenbaum 2004) and purpose of using the data as well as engaging individuals in understanding how the data are used and value created. After establishing new governing principles, new regulations are needed that monitor and punish misuse (rather than deny use) and permit innovative, technology-driven solutions allowing "permissions to flow with the data and ensuring accountability at scale" (Kalapesi 2013, p. 3). In addition, focusing on data usage avoids problems associated with anonymization, which has been shown to be defeated by the same techniques used to analyze big data (de Montjoye et al. 2014, Goroff 2015, PCAST 2014).

How privacy is defined is important for creating a trust model. The traditional definition of privacy focuses on information access. In contrast, the United Nations (UN) Universal Declaration of Human Rights and Warren & Brandeis (1890) focus on what happens to people as a result of how information is used (Gilbert 2007, Kagal & Abelson 2010). The Declaration stipulates "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation" [UN Gen. Assem. Resolut. 217 A (III) 1948].

The *Journal of Privacy and Confidentiality* was established in 2009 (Abowd et al. 2009) to study this changing privacy landscape. Through this journal, the statistical sciences community was introduced to "differential privacy" (Dwork & Smith 2010). Differential privacy is a randomized algorithmic framework that measures the information loss in releasing the perturbed statistical information without compromising individual data. The term differential comes from the notion that the resulting statistical distribution for the quantity of interest is compared under two scenarios: applying the algorithm to the full database and applying the algorithm to the database minus one observation. The difference, or differential, in this comparison gives the measure of potential information leakage (Dwork & Roth 2010, 2013).

The existence of big data and new data flows has spawned a discussion about the need to develop technology options to preserve privacy, called privacy by design (Gilbert 2007, NRC 2007b). Duncan (2007) introduces this concept and highlights the need for clear, reasonable, and transparent definitions of privacy that incorporate risk assessments as a design parameter in data systems. In building privacy by design into data sets, potential disclosure risks of identifying confidential information are built in through "engineering innovation, managerial commitment, informed cooperation of data subjects, and social controls (legislation, regulation, codes of conduct by professional associations, and response to reactions by the public)" (Duncan 2007, p. 1179). Differential privacy is a step in this direction toward confidential data access under use scenarios for which the randomized algorithms can be built. However, differential privacy may not scale to big data applications, in part because of the stringent definition of privacy (Abowd & Schneider 2011).

Record Linkage

One of the newer challenges fueled by the all data revolution has to do with combining data across data sources, particularly in light of multiple owners of the sources. The statistics literature has a long history of methodology development for linking data (Fellegi & Sunter 1969, NRC 1999). This research was initially developed to aid data combination across pairs of data sets for official statistics reporting but has evolved to link data across multiple data sets in the absence of unique identifiers and with multiple matching patterns (Sadinle 2014, Sadinle & Fienberg 2013, Steorts et al. 2014). Simply stated, the goal of a record linkage is to determine whether a record from one data set has a corresponding record in another data set. As identification disclosure risk became magnified due to technology enhancements and network connectivity (Sweeney 2002), the statistical disclosure limitation framework began to expand to record linkage methodology (Skinner 2008).

Today, record linkage technology must take on a new dimension, namely, data governance and the desire to combine data across organizations operating under different, sometimes contradictory, polices and regulations. The integration of record-level data from the data systems of multiple agencies (owners) has the potential for generating high-quality evidence to be used in research and in the assessment of public policy outcomes. However, when attempting to combine data records from these systems, a number of complex legal issues must be considered, not the least of which is the privacy of persons represented by the data in the systems. Because many overlapping and often inconsistent privacy-related restrictions exist at multiple levels of government, the linkage of data, particularly administrative data records, across public agencies can prove exceedingly difficult.

In recent years, one of the most pronounced areas of development in the technological control of data access has been in the area of privacy-protecting record linkage (PPRL; also referred to variably as data matching, data sharing, or entity resolution). This development is most especially pronounced in the literatures of computer science, medical informatics, as well as statistics, with each field producing multiple literature syntheses and taxonomies of the research area (Churches & Christen 2004, Hall & Fienberg 2010, Kum et al. 2014, Vatsalan et al. 2013, Verykios et al. 2009). Steorts et al. (2014) point out that even with the advances in blocking methods for record linkage, including private blocking, PPRL is still an unsolved problem. This is discussed in detail below.

The requirement of privacy adds a third major challenge to the two traditional major challenges of record linkage: quality and scalability. Data quality varies. Real-world data contain errors or so-called dirty data. Therefore, approximate (e.g., probabilistic) matching and classification techniques are required to achieve accurate linkage (Christen 2006, Hernández & Stolfo 1998). The size of a potential match between record sets is the product of the two data sets. Therefore, the computational complexity of a single match between record sets grows quadratically. With large data sets (not even big data), such growth quickly becomes a performance bottleneck, especially when each comparison utilizes approximate matching algorithms made necessary by the existence of dirty data. Accordingly, significant effort has been put forth in the development of techniques to make record matching more scalable (Baxter et al. 2003, Christen 2012, Christen & Goiser 2007, Herzog et al. 2007). Now, in addition to these two challenges, we are adding the need to consider privacy protection (or more precisely, data access restrictions) at every step of the linkage process. It is in this third area of privacy protection that much development has recently occurred.

Recent and extensive overviews of the techniques are being developed specifically to address the privacy protection requirements of the linkage process (e.g., Vatsalan et al. 2014). In their taxonomy, Vatsalan et al. 2014 produce an analysis that begins with the aspects of the privacy situation and details how each technique attempts to address that specific situation. The portion of their taxonomy specific to privacy includes consideration of three dimensions: the number of parties involved, the adversary model assumed, and the actual techniques used in the PPRL approach.

Proposed solutions in the record linkage privacy context can be categorized as belonging to either a three-party or a two-party protocol. In three-party protocols, a third party (in addition to the two with the data) is employed to conduct the linkage. This third party is evaluated at a certain level of trust. Two-party protocols involve only the data sharers in an attempt to achieve higher-level security by reducing the possibility of any third-party collusion. Such approaches necessarily involve more complex computational approaches than the typical three-party approach. In addition, a proposed solution's privacy context can be characterized as containing two types of adversary. Honest-but-curious adversaries will attempt to find out as much as they can without breaking protocol, whereas malicious adversaries are willing to break protocol and attempt different attacks to access as much data as possible. An analogous framework for high-dimensional data analysis is based on cautious or generous sharing of data (Fienberg & Jin 2012).

Lastly, a given technique can be classified as either exact or approximate. Given that in most real-world scenarios the data being matched are often dirty or noisy, if a single trusted shared-identifier (e.g., verified Social Security number) is not available or is not permitted under the data governance constraints, the scenario will typically involve some form of probabilistic linkage. Accordingly, most contemporary techniques seek to address privacy in probabilistic record linkage, reaching back to the earliest approaches of Fellegi & Sunter (1969), but for different reasons.

The most difficult problems involve privacy-preserving probabilistic record linkage in an environment of potentially malicious adversaries. Protecting data in a potentially malicious environment invariably involves the application of some form(s) of one-way hashing (where a new unique value is generated but information is lost, so that recovering the original value is impossible) and/or encryption (where no information is lost and the original value may be recovered). The issue that arises for most approaches to PPRL is that the most common methods of demographic comparison (e.g., string similarity functions) are quickly rendered inadequate by the usage of hashing/encryption techniques. For example, a slight misspelling of a first name will result in completely different hashes and, therefore, a 100% nonmatch. A number of approaches to rectify this issue have been proposed. However, most proposed methods necessarily involve some level of reduction in matching performance as compared with string similarity functions used on unobscured data. As this literature may be less familiar to the statistical science community and other users of statistics, some additional detail is given below.

A protocol employing a common table of reference strings to which the actual strings in two data sets can be compared is proposed so that edit distances from the actual strings to the reference strings can be computed (Pang & Hansen 2006). The reference strings closest to the actual strings can be encrypted and sent along with the edit distances to be used for match determination (Pang & Hansen 2006). Unfortunately, testing of the process results in a fairly sizable reduction in both recall and precision compared with a reference string similarity measure used on unprotected strings (Bachteler et al. 2010).

A protocol that employs mathematical stenography is proposed. The technique is to use the SparseMap method (Hjaltason 2003) to embed a given string into a Euclidean space already populated with random strings (Scannapieco et al. 2007). The coordinates for a given string are then given as the approximate distances between that string and the random strings. A third party compares the Euclidean distances between the strings to determine a match. Testing shows a markedly better result than the table of reference strings approach, but still not an insignificant difference from a reference string similarity measure used on unprotected strings (Bachteler et al. 2010).

A novel method that takes advantage of the properties of Bloom filters is proposed by Schnell et al. (2009). A Bloom filter is a space-efficient probabilistic data structure represented by a bit array that is used to test whether an element is a member of a set. Strings are stored as bits that represent the key-hashed message authentication code of the string's constituent *n*-grams (a contiguous sequence of *n* items). Bloom filters with similar strings will have a high proportion of the same bits set to 1. Using this knowledge, a string similarity measure, the Dice coefficient, can be used to calculate the ratio of similarity. This method works quite well; in fact, testing shows that the precision-recall curves using 2-grams is nearly identical to the benchmark metric (Bachteler et al. 2010). However, the Bloom filter approach is only applicable to using similarity measures that determine the overall similarity of one set of characters to another (that is, when order does not matter). Because of the nature of the encoding scheme and Bloom filters themselves, there is not a way to use this method with standard distance metrics (Winkler 2006).

Successfully deployed in both research and live agency environments (Schroeder 2012), the approach described here uses a newly invented form of ordered hashing where both *n*-gram– and distance-based similarity functions can be employed without loss of recall or precision. That is, the algorithm allows for the comparison of strings using standard string similarity functions when the strings must be obfuscated in some secure manner (e.g., hash, encrypt).

The approach employs two separate processes. The first uses a one-time pad to create a cipher for each string, providing for a type of obfuscation that is both theoretically unbreakable and not vulnerable to frequency analysis (Denning & Elizabeth 1982). The second process enhances the first by employing the method of chaffing and winnowing, which involves the addition of fake characters (the chaff) to the valid characters (the wheat) so as to result in all encoded strings being the same length (Rivest 1998). The proposed methodology is currently being deployed in two multiagency data-integration projects in Virginia (Schneider et al. 2012, Spears et al. 2012).

One issue still in need of further study is the propagation of errors invoked in the linkage process to subsequent linkages (Sadinle & Fienberg 2013) and data analyses (Harron et al. 2014, Kim & Chambers 2012). Specification of any additional effect of the ordered hashing process on error propagation remains an open question. Also, scalability of the method to close-to-real-time big data settings is questionable. In the current application context, the system routinely carries out the joining of data sets of more than 30 million rows. However, the processing time for a very large merge can easily be 3 or 4 hours. As the current primary use of the system is the production of longitudinal data sets to be used in program and policy analyses, these times are not an issue. However, if the intended use of the system changes and a requirement for close-to-real-time linkage arises, the current implementation of the algorithm would most likely need to be revisited, potentially with an eye toward improved blocking method performance (Steorts et al. 2014).

This research on privacy-preserving data systems that are able to integrate data across multiple sources governed by different policies opens more opportunities for gaining interoperability between data assets that could be used productively for research. These systems also offer the opportunity for monitoring use through automated audits or other mechanisms. The potential benefits of moving to a new system that controls use of data and not the collection of the data is expected to yield large societal benefits (Kalapesi 2013, PCAST 2014). Challenges to implementing these processes are creating the political will and public understanding of such approaches (Acquisti et al. 2015, Kagal & Abelson 2010), the creation of significant criminal penalties for privacy violations, and figuring out practical ways for individuals to express their preferences about personal data. One suggestion is to allow individuals to delegate their choices to organizations they trust (Mundie 2014). Education and technology are also important for changing perceptions and increasing understanding about the proposed new approach, especially for older generations that may perceive a trust-based approach as threatening (PCAST 2014).

Moving to a Trust-Centric Approach

Although our focus has been on the role of privacy in the federal statistical system and changes occurring due to the all data revolution, much can be learned from examining changes in other fields, especially those also grappling with big data and privacy issues. Adoption of the principles set forth in the World Economic Forum (Kalapesi 2013), the President's Council of Advisors on Science and Technology (PCAST 2014), and other reports proposed for genomic research (Erlich et al. 2014) is a first step in focusing on a trust not privacy approach (Erlich et al. 2014). Genomic research requires analysis of massive data sets, and current privacy models rely on de-identification techniques referred to as zero-sum choices. To create a win-win situation where both researchers and participants benefit, Erlich et al. (2014) propose three trust-based principles that reward good behavior, deter malicious behavior, and punish noncompliance:

- Transparency creates trust among parties, implemented by informing participants about potential and actual uses of the data. This is similar to current principles set forth in the United States, in Canada, and by the OECD.
- Increased control over future data use creates trust.
- Reciprocity maintains trust when mechanisms are created that allow participants to reward researchers who increase scientific knowledge and punish those who violate trust.

Similar to the current sharing services, such as ride sharing and home sharing, a bilateral consent framework is proposed. The implementation of this framework requires (*a*) the creation of a participatory community with participants, researchers, and trusted mediators; (*b*) an audit system; and (*c*) a system to establish reputation based on participant ratings, peer researcher recommendations, prior studies, the integrity of the host organization, and audit reports (Erlich et al. 2014). The proposed approach relies on technology but is technology-agnostic, is flexible and adaptable, and allows for choices and decisions to be made based on context. Ideally, this trust-based approach would become self-monitoring. Although this is an exciting emerging theme in the literature, it is likely that researchers who propose differential privacy approaches would strongly disagree that a trust-based approach could be adopted on a wide enough scale and be monitored.

The tension in using big data is the expectation (and hype) that social benefits will outweigh the privacy risks of using these data, yet the culmination of small biases has the potential to affect outcomes of disadvantaged groups. At the same time, individuals and groups of individuals can use these new sources of data and methods to ensure their rights. The trust-centric approach may overcome the concern that big data analytics may lead to inequitable treatment of disadvantaged groups or create such an opaque decision-making environment that individual autonomy is lost in an impenetrable set of algorithms (EOP 2014).

CONCLUSIONS

The all data revolution is changing the current privacy paradigm from statistical disclosure limitation, essentially concealment, to a trust-but-verify approach, but legislation has not changed to reflect this. Theory and frameworks have been developed and are evolving to meet this new paradigm, and remote access infrastructure is in place to allow new approaches to emerge. Statistical research is needed to build on existing risk–utility frameworks while accommodating the growing amounts of nonstatistically designed data. There will always be a race between those who want to conceal and those who want to uncover data. Privacy solutions have to clearly outline parameters and boundaries. As new sources of data, technology, and culture evolve, research must also evolve to accommodate the changing privacy paradigm.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This research was partially based upon work supported by the National Science Foundation under Grant No. 1338491. We would like to thank Daniel H. Weinberg for his review of the paper.

LITERATURE CITED

Abowd JM, Lane JI. 2003. *Synthetic data and confidentiality protection*. Tech. Pap. TP-2003-10, Cent. Econ. Stud., US Census Bur., Washington, DC

Abowd JM, Lane JI. 2004. New approaches to confidentiality protection: synthetic data, remote access and research data centers. In *Privacy in Statistical Databases*, ed. J Domingo-Ferrer, V Torra, pp. 282–89. Lect. Notes Comput. Sci. Ser. 3050. Berlin: Springer

Abowd JM, Nissim K, Skinner CJ. 2009. First issue editorial. *J. Priv. Confid.* 1(1):1

Abowd JM, Schneider MJ. 2011. An application of differentially private linear mixed modeling. *IEEE 11th Int. Conf. Data Min. Workshops*, ed. M Spiliopoulou, H Wang, D Cook, J Pei, W Wang et al., pp. 614–19. New York: IEEE

Abowd JM, Stephens BE, Vilhuber L, Andersson F, McKinney KL, et al. 2009. The LEHD infrastructure files and the creation of the quarterly workforce indicators. In *Producer Dynamics: New Evidence from Micro Data*, ed. T Dunne, JB Jensen, MJ Roberts, pp. 149–230. Chicago: Univ. Chicago Press

Acquisti A, Brandimarte L, Loewenstein G. 2015. Privacy and human behavior in the age of information. *Science* 347(6221):509–14

Agafiţei M, Gras F, Kloek W, Reis F. 2015. Measuring output quality for multisource statistics in official statistics: some directions. *Stat. J. IAOS* 31(2):203–11

Bachteler T, Schnell R, Reiher J. 2010. An empirical comparison of approaches to approximate string matching in private record linkage. *Proc. Stat. Can. Symp. 2010: Social Statistics: The Interplay among Censuses, Surveys and Administrative Data*, Ottawa, Can., pp. 290–95. Ottawa, Can.: Stat. Can.

Baxter R, Christen P, Churches T. 2003. A comparison of fast blocking methods for record linkage. *Proc. 2003 ACM SIGKDD Workshop Data Clean., Rec. Link. Object Consol.*, pp. 25–27. New York: ACM

Berinsky AJ, Huber GA, Lenz GS. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Pol. Anal.* 20(3):351–68

Bohannon J. 2011. Social science for pennies. *Science* 334(6054):307

Bohannon J. 2015. Privacy. Credit card study blows holes in anonymity. *Science* 347(6221):468

Braaksma B, Zeelenberg K. 2015 "Re-make/re-model": Should big data change the modelling paradigm in official statistics? *Stat. J. IAOS* 31:193–202

Brennan N, Conway PH, Tavenner M. 2014. The Medicare physician-data release-context and rationale. *N. Engl. J. Med.* 371(2):99–101

Burwell S. 2014. *Guidance for providing and using administrative data for statistical purposes.* Memo. OMB M-14-06, Off. Manag. Budg., Washington, DC

Campbell S, Shipp S, Mulcahy T, Allen T. 2009. Informing public policy on science and innovation: the Advanced Technology Program's experience. *J. Technol. Transf.* 34(3):304–19

CDC (Cent. Dis. Control Prev.). 2015a. *How NCHS protects your privacy*. Cent. Dis. Control Prev., Atlanta. **http://www.cdc.gov/nchs/about/policy/confidentiality.htm**

CDC (Cent. Dis. Control Prev.). 2015b. *NCHS Research Data Center (RDC)*. Cent. Dis. Control Prev., Atlanta. **http://www.cdc.gov/rdc/index.htm**

Christen P. 2006. A comparison of personal name matching: techniques and practical issues. *Sixth IEEE Int. Conf. Data Min. Workshops*, Hong Kong, pp. 290–94. New York: IEEE

Christen P. 2012. *The Data Matching Process*. Berlin: Springer

Christen P, Goiser K. 2007. Quality and complexity measures for data linkage and deduplication. In *Quality Measures in Data Mining*, ed. F Guillet, HJ Hamilton, pp. 127–51. Berlin: Springer

Churches T, Christen P. 2004. Some methods for blindfolded record linkage. *BMC Med. Inform. Decis. Mak.* 4(1):9

Cox LH. 1982. Solving confidentiality protection problems in tabulations using network optimization: a network model for cell suppression in the US economic censuses. *Proc. Int. Semin. Stat. Confid.*, Dublin, pp. 229–45. The Hague, Neth.: Int. Stat. Inst.

Dalenius T, Reiss SP. 1982. Data-swapping: a technique for disclosure control. *J. Stat. Plan. Inference* 6(1):73–85

de Montjoye Y-A, Radaelli L, Singh VK, Pentland AS. 2014. Unique in the shopping mall: on the re-identifiability of credit card metadata. *Science* 347(6221):536–39

Denning DE. 1982. *Cryptography and Data Security*. Reading, MA: Addison-Wesley

Duncan G. 2007. Privacy by design. *Science* 317(5842):1178–79

Duncan G, Fienberg S. 1997. Obtaining information while preserving privacy: a Markov perturbation method for tabular data. *Jt. Stat. Meet. Proc.*, pp. 351–62. Alexandria, VA: Am. Stat. Assoc.

Duncan G, Keller-McNulty S, Stokes SL. 2001. *Disclosure risk versus data utility: the RU confidentiality map*. Tech. Rep. 121, Natl. Inst. Stat. Sci., Research Triangle Park, NC

Duncan G, Keller-McNulty S, Stokes S. 2004. *Database security and confidentiality: examining disclosure risk versus data utility through the RU confidentiality map*. Tech. Rep. 142, Natl. Inst. Stat. Sci., Research Triangle Park, NC

Duncan G, Lambert D. 1986. Disclosure-limited data dissemination. *J. Am. Stat. Assoc.* 81(393):10–18

Duncan G, Lambert D. 1989. The risk of disclosure for microdata. *J. Bus. Econ. Stat.* 7(2):207–17

Dwork C, Roth A. 2013. The algorithmic foundations of differential privacy. *Theor. Comput. Sci.* 9(3–4):211–407

Dwork C, Smith A. 2010. Differential privacy for statistics: what we know and what we want to learn. *J. Priv. Confid.* 1(2):135–54

EOP (Exec. Off. Pres.). 2014. *Big data: seizing opportunities, preserving values*. Rep., Exec. Off. Pres., Washington, DC. **https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf**

Erlich Y, Williams JB, Glazer D, Yocum K, Farahany N, et al. 2014. Redefining genomic privacy: trust and empowerment. *PLOS Biol.* 12(11):e1001983

FCC (Fed. Commun. Comm.). 2011. *Annual report and analysis of competitive market conditions with respect to mobile wireless, including commercial mobile services*. Rep., Fed. Commun. Comm., Washington, DC. **https://www.fcc.gov/document/15th-mobile-wireless-competition-report**

Fellegi IP, Sunter AB. 1969. A theory for record linkage. *J. Am. Stat. Assoc.* 64(328):1183–210

Fienberg SE, Jin J. 2012. Privacy-preserving data sharing in high dimensional regression and classification settings. *J. Priv. Confid.* 4(1):221–43

Gilbert N. 2007. Dilemmas of privacy and surveillance: challenges of technological change. *Crim. Justice Matters* 68(1):41–42

Gomatam S, Karr A, Reiter J, Sanil A. 2005. Data dissemination and disclosure limitation in a world without microdata: a risk–utility framework for remote access analysis servers. *Stat. Sci.* 20(2):163–77

Goroff DL. 2015. Balancing privacy versus accuracy in research protocols. *Science* 347(6221):479–80

Hall R, Fienberg SE. 2010. Privacy-preserving record linkage. In *Privacy in Statistical Databases*, ed. J Domingo-Ferrer, E Magkos, pp. 269–83. Lect. Notes Comput. Sci. Ser. 6344. Berlin: Springer

Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. 2014. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med. Res. Methodol.* 14:36

Hernández MA, Stolfo SJ. 1998. Real-world data is dirty: data cleansing and the merge/purge problem. *Data Min. Knowl. Discov.* 2(1):9–37

Herzog TN, Scheuren FJ, Winkler WE. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer

Hjaltason Gísli R.Hanan Samet. 2003. Properties of embedding methods for similarity searching in metric spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* 25(5):530–49

IOM (Inst. of Med.). 2009. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Washington, DC: Natl. Acad. Press

Kagal L, Abelson H. 2010. Access control is an inadequate framework for privacy protection. *W3C Workshop Priv. Adv. Web APIs*, London, pp. 1–6. **http://www.w3.org/2010/api-privacy-ws/papers/privacy-ws-23.pdf**

Kalapesi C. 2013. *Unlocking the value of personal data: from collection to usage*. Tech. Rep., World Econ. Forum, Geneva. **www3.weforum.org/. . ./WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf**

Karr A, Reiter JP. 2014. Analytical frameworks for data release: a statistical view. In *Privacy, Big Data and the Public Good*, ed. J. Lane, V Stodden, S Bendor, H Nissenbaum, pp. 276–95. New York: Cambridge Univ. Press

Keller S, Koonin S, Shipp S. 2012. Big data and city living—What can it do for us? *Significance* 9(4):4–7

Keller S, Shipp S. 2016. Building resilient cities: harnessing the power of urban analytics. In *The Resilience Challenge: Looking at Resilience through Multiple Lenses*. Springfield, IL: Thomas. In press

Keller-McNulty S, Nakhleh C, Singpurwalla N. 2005. A paradigm for masking (camouflaging) information. *Int. Stat. Rev.* 73(3):331–49

Keller-McNulty S, Unger E. 1993. Database systems: inferential security. *J. Off. Stat.* 9:475–99

Keller-McNulty S, Unger E. 1998. A database system prototype for remote access to information based on confidential data. *J. Off. Stat.* 14:347–60

Kim G, Chambers R. 2012. Regression analysis under probabilistic multi-linkage. *Stat. Neerl.* 66(1):64–79

Kum H-C, Krishnamurthy A, Machanavajjhala A, Reiter MK, Ahalt S. 2014. Privacy preserving interactive record linkage (PPIRL). *J. Am. Med. Inform. Assoc.* 21(2):212–20

Landau S. 2015. Control use of data to protect privacy. *Science* 347(6221):504–6

Lane J, Shipp S. 2008. Using a remote access data enclave for data dissemination. *Int. J. Digit. Curation* 2(1):128–34

Lazer DM, Kennedy R, King G, Vespignani A. 2014. The parable of Google Flu: traps in big data analysis. *Science* 14:1203–5

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, et al. 2011. *Big data: the next frontier for innovation, competition, and productivity*. Tech. Rep., McKenzie Glob. Inst., San Francisco. **http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation**

Mason W, Suri S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* 44(1):1–23

Mundie C. 2014. Privacy pragmatism. *Foreign Aff.* 93(2):7–8

Nissenbaum H. 2004. Privacy in context: technology, policy, and the integrity of social life. *Wash. Law Rev.* 79(1):119–57

NRC (Natl. Res. Counc.). 1993. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 1999. *Record Linkage Techniques—1997: Proc. Int. Workshop Expo*, Arlington, VA. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2007a. *Engaging Privacy and Information Technology in a Digital Age*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2007b. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2013. *Frontiers in Massive Data Analysis*. Washington, DC: Natl. Acad. Press

OECD (Organ. Econ. Co-op. Dev.). 1980. *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. Paris: Organ. Econ. Co-op. Dev.

OECD (Organ. Econ. Co-op. Dev.). 2013. *The OECD Privacy Framework*. Paris: Organ. Econ. Co-op. Dev.

OMB (Off. Manag. Budg.). 2007. 72 FR 33361—Implementation guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). *Fed. Regist.* 72, no. 115 (June 15, 2007): 33361–77. Washington, DC: US Gov. Publ. Off. **https://www.gpo.gov/fdsys/pkg/FR-2007-06-15/pdf/E7-11542.pdf**

Pang C, Hansen D. 2006. Improved record linkage for encrypted identifying data. *Proc. 14th Annu. Health Inform. Conf.*, Sydney, pp. 164–68. Brunswick East, Aust.: Health Inform. Soc. Aust.

PCAST (Pres. Counc. Advis. Sci. Technol.). 2014. *Big data and privacy: a technology perspective*. Rep. Exec. Off. Pres. Pres. Counc. Advis. Sci. Technol., Washington, DC. **https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf**

Raghunathan TE, Reiter JP, Rubin DB. 2003. Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* 19(1):1–16

Reiter JP. 2005a. Estimating risks of identification disclosure in microdata. *J. Am. Stat. Assoc.* 100(472):1103–12

Reiter JP. 2005b. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *J. R. Stat. Soc. A* 168(1):185–205

Reiter JP. 2009. Multiple imputation for disclosure limitation: future research challenges. *J. Priv. Confid.* 1(2):223–33

Reiter JP, Wang Q, Zhang B. 2014. Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *J. Priv. Confid.* 6(1):17–33

Rivest RL. 1998. Chaffing and winnowing: confidentiality without encryption. *CryptoBytes* (*RSA Lab.*) 4(1):12–17

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley

Rubin DB. 1996. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* 91(434):473–89

Sadinle M, Fienberg SE. 2013. A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. *J. Am. Stat. Assoc.* 108(502):385–97

Sadinle M. 2014. Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Ann. Appl. Stat.* 8(4):2404–34

Scannapieco M, Figotin I, Bertino E, Elmagarmid AK. 2007. Privacy preserving schema and data matching. *Proc. ACM SIGMOD Int. Conf. Manag. Data*, Beijing, pp. 653–64. New York: ACM

Schneider MJ, Abowd JM. 2015. A new method for protecting interrelated time series with Bayesian prior distributions and synthetic data. *J. R. Stat. Soc. A* 178(4):963–75

Schneider MJ, Massa T, Vivari B. 2012. *The earning power of recent graduates from Virginia's colleges and universities: How are graduates from different degree programs doing in the labor market?* Rep., Econ. Success Metr. Proj., Am. Inst. Res., Washington, DC. **http://www.air.org/sites/default/files/downloads/report/Virginia_EMS_Report1_0.pdf**

Schnell R, Bachteler T, Reiher J. 2009. Privacy-preserving record linkage using Bloom filters. *BMC Med. Inform. Decis. Mak.* 9(1):41

Schouten B, Cigrang M. 2003. Remote access systems for statistical analysis of microdata. *Stat. Comput.* 13(4):381–89

Schroeder AD. 2012. Pad and chaff: secure approximate string matching in private record linkage. *Proc. 14th Int. Conf. Inform. Integr. Web-Based Appl. Serv.*, Bali, Indones., pp.121–25. New York: ACM

Schwab K, Marcus A, Oyola J, Hoffman W, Luzi M. 2011. *Personal data: the emergence of a new asset class*. Tech. Rep., World Econ. Forum, Geneva. **http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf**

Skinner CJ. 2008. Assessing disclosure risk for record linkage. In *Privacy in Statistical Databases*, ed. J Domingo-Ferrer, Y Saygin , pp. 166–76. Lect. Notes Comput. Sci. Ser. 5262. Berlin: Springer

Spears JV, Bradburn I, Schroeder AD, Tester D, Forry N. 2012. New data on child care subsidy programs. *Policy Pract.* 2012(Aug.):18–21

Steorts RC, Hall R, Fienberg SE. 2014. SMERED: a Bayesian approach to graphical record linkage and de-duplication. arXiv:1403.0211 [stat.CO]

Sweeney L. 2002. k-Anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10(5):557–70

UN Gen. Assem. Resolut. 217 A (III). 1948. *Universal Declaration of Human Rights*, Dec. 10. UN Doc. A/810. United Nations, New York. **http://www.un.org/en/documents/udhr/index.shtml**

US Census Bur. 2013. *Center for Economic Studies and Research Data Centers research report: 2012*. Rep. Res. Methodol. Dir., US Census Bur., Washington, DC

US Census Bur. 2015a. *Data protection and privacy. Data protection.* US Census Bur., Washington, DC. **https://www.census.gov/about/policies/privacy/data_protection.html**

US Census Bur. 2015b. *Data protection and privacy. Title 13—Protection of confidential information.* US Census Bur., Washington, DC. **http://www.census.gov/about/policies/privacy/data_protection/title_13_-_protection_of_confidential_information.html**

US Census Bur. 2015c. *Privacy and confidentiality. Title 26, US Code.* US Census Bur., Washington, DC. **https://www.census.gov/history/www/reference/privacy_confidentiality/title_26_us_code_1.html**

US Census Bur. 2015d. *History of public use microdata areas (PUMAs): 1960–2000.* U.S. Census Bur., Washington, DC

US DOE (US Dep. Educ.). 2015. *Confidentiality laws*. Stat. Stand. Program, Natl. Cent. Edu. Stat., US Dep. Educ., Washington, DC. **http://nces.ed.gov/statprog/conflaws.asp**

US DOJ (US Dep. Justice). 2015. *What* is *FOIA?* US Dep. Justice, Washington, DC. **http://www.foia.gov/index.html**

VASEM (Virg. Summit Sci. Eng. Med.). 2014. *Meeting on big data: report of December 4–5, 2014*. Rep., Virg. Summit Sci. Eng. Med., Washington, DC. **http://seas.virginia.edu/admin/vasem/news/pdfs/vasem_big_data_2014.pdf**

Vatsalan D, Christen P, O'Keefe CM, Verykios VS. 2014. An evaluation framework for privacy-preserving record linkage. *J. Priv. Confid.* 6(1):35–75

Vatsalan D, Christen P, Verykios VS. 2013. A taxonomy of privacy-preserving record linkage techniques. *Inform. Syst.* 38(6):946–69

Verykios VS, Karakasidis A, Mitrogiannis VK. 2009. Privacy preserving record linkage approaches. *Int. J. Data Min. Model. Manag.* 1(2):206–21

Wallman KK, Harris-Kojetin BA. 2004. Implementing the Confidential Information Protection and Statistical Efficiency Act of 2002. *Chance* 17(3):21–25

Warren S, Brandeis L. 1890. The right to privacy. *Harvard Law Rev.* 4:193–220

Winkler WE. 2006. *Overview of record linkage and current research directions*. Res. Rep., Stat. 2006-2. Stat. Res. Div., US Census Bur., Washington, DC