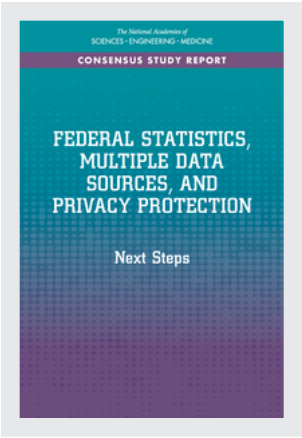

Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps (2017)

DETAILS

194 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-46537-3 | DOI 10.17226/24893

CONTRIBUTORS

Robert M. Groves and Brian A. Harris-Kojetin, Editors; Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2017. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24893.

5

Preserving Privacy Using Technology from Computer Science, Statistical Methods, and Administrative Procedures

In this chapter we examine how statistical agencies implement the legal requirements described in Chapter 4 to protect the privacy of the data they collect or acquire. We also build upon the discussion from our first report regarding restricting data, restricting access, and using computer science and cryptography to protect the privacy of statistical data in the context of using multiple data sources (National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 5). We begin by drawing a distinction between two avenues of privacy breach in the context of statistical data analysis: (1) threats to the security of the raw data, and (2) the use of statistical findings, aggregations, and conclusions drawn from the confidential data to learn about an individual or organization. As in the previous chapter, we focus only on individual privacy breaches to the data providers: that is, breaches that could not have occurred if the individual's or the organization's data had not been in the confidential dataset. Inferences based on facts about the population as a whole are not viewed as individual privacy breaches.

For each of the two avenues, we describe specific vulnerabilities and some of the technologies that address or mitigate these risks. We note some of the costs of privacy protection and discuss approaches for statistical agencies to begin to evaluate new technologies and methods for protecting privacy. It is important to note that the technologies and methods have different purposes in protecting data. In addition, some techniques have been used for a longer time in statistical agencies than others so their uses are at least partly understood; others are not in widespread use.

TWO AVENUES TO A BREACH OF PRIVACY

There are two distinct avenues to an individual privacy breach: security threats and inferential disclosure. Security threats comprise most usual concerns and occur in many ways. Most simply, there may be a data breach (e.g., someone breaks in or eavesdrops) or data loss (e.g., a laptop containing data is left on a train). Other, more deliberate, ways can involve data integrity problems (e.g., an outsider or an insider has tampered with the data, deleted it, or added spurious records) or bribery (e.g., an insider with authorized access sells data access or sells a copy of the data). There are also a variety of side-channel attacks, such as determining which data are accessed, and potentially the values of these data, based on physical measurements—acoustical properties, power consumption—of the computing device and the time used to carry out certain steps of the computation. All of these are threats to the raw data, the individual records.

The second avenue to privacy breach is more subtle. As noted in Chapter 4, this avenue involves being able to learn something about an individual in a dataset. This kind of breach is often referred to as *inferential disclosure* in the statistics literature, but this term is sometimes used in a very different sense. To avoid confusion, we use the term *individual privacy breach*. The concept is easily explained through an example. Suppose statistical analysis of a medical dataset shows that smoking causes cancer. The scientific conclusion of the study allows one to infer that any smoker is at increased risk of cancer. This is true whether or not the data of a particular smoker are in the medical dataset. So far, this is not an individual privacy breach—learning that smoking causes cancer was the point of the study. An individual privacy breach occurs when one can learn something about a participant in the study that cannot be learned about someone not in the study.

For example, if the holders of the medical data were to release poorly "anonymized" individual medical records of the study participants and if it was possible to determine from the records, say, information about the date in which someone entered the study, it might be possible to learn specific details about the person's medical history by tracing individual records back to the corresponding participants or through more mathematically subtle means. Surprisingly, individual privacy breaches can occur even when only simple statistics about the study participants are made public. Attacks on individual privacy can occur by combining the legitimate and correct outputs of a statistical analysis with auxiliary information from other sources—such as other studies (including replications), information people publish on their websites, previous measurements for the same area or population, product recommendations, blogs, social networking sites, and information one has about colleagues and peers—to learn facts about individual data providers that could not have been deduced had the

individual not contributed data. In other words, *by definition*, only data providers are at risk of inferential threats of privacy breaches. It is important to note that these individual privacy threats persist even if all security threats are eliminated.

Security Threats

Securing Data

Many security threats can be addressed by maintaining data in a secure form—partitioned or encrypted—which includes protection against data loss. Data loss can occur for many reasons. One is due to physical loss of the storage medium because of fire or flooding. To mitigate this danger, all data of value need to be backed up or replicated. Another reason for data loss is due to tampering, such as an unauthorized update being applied. Replication may be helpful in this case because an intruder is forced to tamper with multiple copies to install the update. However, computer security methods are much more likely to prevent tampering and intrusions in the first place.

Security threats are also addressed by access protocols and technologies, such as multifactor authentication, to prevent unauthorized access to data. As described in Chapter 4, datasets may be protected by different laws, and access needs to be provided to only those people authorized to use the data, who are typically sworn under penalty of law not to disclose the information or use it for anything other than a statistical purpose. Administrative procedures, such as access lists, are used to ensure that only authorized users are permitted to log into particular sites and access particular data repositories. That is, specific users or Internet protocol addresses are safelisted (if allowed access) or blacklisted (if forbidden). Alternatively, rather than authorizing users, roles can be authorized, that is, access is controlled through a role-based access-control paradigm. In this approach, users having a given role are provided with access authorized for that particular role. However, it might be necessary to add a role-based context-constrained access control to guarantee that external constraints are additionally imposed on the data access.[1] That is, while a role-based access-control paradigm provides capabilities based on roles, a role-based context-constrained access control accounts for additional context constraints, potentially externally extracted, imposed on the given role. For example, a public employee (a role) may be able to access a range of sensitive government information about individuals to perform her job, but she

[1]This constraint is sometimes referred to as a context primitive.

would be prohibited from accessing such information on family members (contextual constraint).

Unfortunately, data breaches still may occur, and so efforts have to be made to minimize their effects. Often, these breaches could be due to guessing or stealing a password from an authorized user rather than through a successful attack on a system's defenses. As such, systems need to be designed to minimize the harm caused by any breach. Best practices in this regard include storing the data in encrypted form, so that an additional decryption step has to be performed by anyone who manages to get unauthorized access to the encrypted data. One such practice is storing the data in distinct partitions so that any one breach exposes only one partition rather than all the data. Another practice is breaking data into "shares" and storing the shares separately, so that reconstructing even a single data element requires obtaining most or even all of the shares.

Ideally, data are partitioned so that what is in any one partition is not particularly sensitive even if a combination of partitions contains very sensitive data. A simple example is to have an identification (ID) to name and address mapping in one partition and an ID to income mapping in a second partition. Only by putting the two partitions together could an intruder learn the mapping between names and incomes.

In the case of statistics generated from multiple sources, it is possible that joined data across sources are far more sensitive than data in any one source. But the joined data may only be an ephemeral intermediate result, which is aggregated down to statistical products that can be made public. In such situations, it may be desirable to avoid ever storing the full granular joined data. Technology for secure multiparty computing could, in some situations, even permit a statistical agency to compute the desired aggregate result without ever actually learning all the detailed data in each of the data sources.

Securing Computation

As we describe in Chapter 3, statistical analyses using multiple data sources can be accomplished in a wide variety of computing architectures. All (and, if possible, only) the data required for a given analysis or task could be copied to a single location ("centralized") and retained only while needed for the statistical computations. This is an example of the general concept of principle of data minimization: security risks are mitigated both by constraining the scope of the collected information and the duration for which it is held. Indefinitely housing multiple data sources together exacerbates the risks of disclosure—there is more time for an adversary to find a successful attack—and it also presents a richer, and therefore more tempting, target.

At the opposite end of the spectrum, computation on multiple data sources can be decentralized: the data never leave their individual original locations, and the multiple locations engage in a cryptographic protocol to cooperatively compute the outcome of the statistical computations.[2] The protocol ensures that no party to the analysis (i.e., none of the individual sources) learns about the data held by the other parties, other than what can be inferred from the output and the party's own input. This kind of secrecy guarantee is easily illustrated with a simple example: suppose Alice and Bob wish to determine who read more books in the previous week, but they don't wish to reveal how many books they read. Suppose further that both will engage in a cryptographic protocol for achieving this comparison without cheating—they are honest but curious (Goldreich, 2004). Say that Alice read seven books, and Bob read five. The output of the protocol simply says "Alice," since Alice read more books than Bob. From this output, Bob can infer a little extra: he knows that Alice read at least six books, but he cannot determine whether Alice read seven books (reality) or 17 books. Bob has learned no more than what can be determined from the output of the protocol ("Alice") and his own input (five). This example also shows that what Bob can infer about Alice from the output of the protocol depends also on Bob's own inputs.

Cryptographic protocols of this type are instances of secure multiparty computation, also known as secure function evaluation. To study or implement secure multiparty computation, one needs to consider three factors:

1. What types of bad behavior does the protocol need to protect against? Are the parties honest but curious? Might the parties deviate from the protocol, either deliberately or if the host machine is invaded by a virus?
2. What fraction of the parties would need to collude to circumvent the protections of the protocol?
3. How many (total) parties are participating in the protocol?

The past decade has seen tremendous progress in secure multiparty computation for the two-party, honest but curious (also known as semi-honest) case (see, e.g., the optimizations noted in Gueron et al., 2015). Secure versions of computations whose nonprivate versions require upwards of a trillion Boolean gates (the simplest computational element in a digital computer) are now routine. In the future, it may become practical for a physical

[2]A cryptographic protocol permits two or more parties to cooperatively make decisions based on, or learn specifics about, the combination of their individual information, without explicitly revealing their information to one another.

data warehouse to be replaced by a virtual data warehouse through secure multiparty computation and other advanced cryptographic technology.

Whether in a centralized setting or at each individual data source, outsider threats should be reduced by encrypting data both at rest and in transit. Other cryptographic techniques are also available for ensuring that stored data have not been tampered with (memory checking).

In the future, it is expected that advanced cryptographic techniques, for which proofs of concept already exist, will permit computation on encrypted data without the need to decrypt (Gentry, 2009; also see Canetti et al., 2017). Such technology can mitigate many of the concerns with centralization without the complexity of secure multiparty computation. In a related vein, proofs of concept also exist for encrypting data so as to permit only the results of certain prespecified computations to be decrypted without a special key (Boneh et al., 2010). Although they are not yet mature, these emerging technologies would be powerful enablers of securely combining multiple data sources in both centralized and decentralized settings. Using current statistical techniques, analysts often examine data, for example, to identify outliers and data errors. If and when these processes can be described in algorithmic terms, they too could be carried out using advanced cryptographic techniques. Other analytical goals, such as assessing whether or not there are sufficient numbers of cases of interest to support conclusions, can already be achieved using current cryptographic techniques.

Modern cryptography is founded on the principle that certain kinds of computational problems are too complex to be carried out in practice. For example, encryption systems require fast algorithms for encrypting data and fast algorithms for decrypting encrypted messages with the help of the secret decryption key, but the systems also require that there is no practical algorithm that would allow someone intercepting the encrypted message to decrypt it *without* the secret key. In other words, decrypting (without the secret decryption key) is computationally hard. The famous RSA cryptosystem,[3] for example, rests on the assumption that factoring a number, $n = pq$, where p and q are large primes, is computationally hard.

In cryptography, algorithms (e.g., for encryption) are often modeled as "black boxes" whose internal states and inputs—such as secret keys and (intentionally) random bits—are not visible to an adversary. Algorithms can then be proven to be secure under various assumptions about the hardness of certain computations, such as factoring. This is not the end of the story, however. The algorithms are often run in poorly secured settings, and "side-channel attacks" can exploit the physical properties of the

[3]A system in which the encryption key is public and different from the decryption key, which is kept private. It is named after the three authors who first described it.

devices on which the cryptographic algorithms are running. The physical implementation might enable "leakage" of the internal state of the algorithm, which is defined as observations and measurements of the internal characteristics of a cryptographic algorithm, including random choices made by the algorithm and secret keys, such as a decryption key for an encryption algorithm. For example, the RSA encryption algorithm requires a series of operations that depend on the individual bits in the secret key, but the operation associated with "0" is faster than the operation associated with "1." There are also differences in the levels of power consumed by the two operations. Attacks exploiting these differences can and have broken systems with a mathematical security proof without violating any of the underlying mathematical principles; for timing attacks, see Kocher (1996); for differential power analysis, see Kocher et al. (1998); for acoustic cryptanalysis, see Genkin et al. (2013). However, there is leakage-resilient cryptography that defends against such attacks; see Akavia et al. (2009) and Goldwasser and Rothblum (2010).

Other security threats arise if algorithms are run on untrusted servers or in the cloud. For example, one may have a large amount of sensitive data, such as genomic data, stored on the untrusted server. Even if the data are encrypted, the server can observe data access patterns, potentially allowing the server to infer which medical test is being applied to the genomic data. Or, consider a program that tests a given location in the DNA and branches according to whether the value is a "C" or a "T." That is, the program examines the data to see if it is a "C" or a "T" and, according to the output, proceeds with one of two possible computations. Suppose further that these two computations access data stored in different parts of the computer memory. Then, an untrusted server may be able to infer from the program's data access pattern which of the two possible computations was carried out, allowing inference of the correct protein, "C" or "T." These kinds of threats are addressed in the oblivious random access machine literature (Goldreich et al., 1987; Wang et al., 2015).

Even if a server is trusted, for example, in certain cloud computing environments, information can pass from one application to another running on the same server in what is known as a cross-VM (virtual machine) attack. A VM operates as a complete computer system, providing the complete functionality of a computer. Virtualization permits unrelated programs to be run simultaneously on a single computer, allowing multiple computing environments to behave as if each is run in isolation. Services, such as Microsoft's Azure, Google's Compute Engine, and Amazon's EC2, allow users to instantiate VMs on demand. As a result, VMs instantiated by distinct cloud customers share a physical infrastructure. In a cross-VM attack, a malicious program on one VM can learn information about co-resident instances. For example, (time-shared) caches allow an attacker to measure

when other instances are experiencing computational load. Leaking such information is less innocuous than it appears, permitting "co-residence detection (agnostic to network configuration), surreptitious detection of the rate of web traffic a co-resident site receives, and even timing keystrokes by an honest user (via SSH) of a co-resident instance" (Ristenpart et al., 2009, p. 9).

Risks from "inside jobs" are also problematic, although these too can be mitigated, for example, with untamperable and auditable logs and possibly with cryptographic copyright protection methods to discourage selling the data or access to the data.

Of the two classes of risks—security and individual privacy—the former is better understood. Altman and colleagues (2015, p. 28) note that the formal guidance agencies are given for analyzing and mitigating related information security risks is "voluminous, proscriptive, specific, actionable, frequently updated, and integrative into systems of legal audit and certification." In contrast, the guidance for identifying and mitigating individual privacy risks is "general, abstract, infrequently updated, and self-directed," which can often lead to "inconsistent identification of privacy risks and ineffective application of privacy safeguards" (p. 28).

Modern cryptography has studied aspects of security problems in detail, starting in the late 1970s (see Diffie and Hellman, 1976; Rivest et al., 1978) with a rigorous theory beginning in the early 1980s (Goldwasser and Micali, 1982; Goldwasser et al., 1988). The science of inferential disclosure risk dates back to the 1970s (see Fellegi, 1972), but a rigorous theory of privacy loss began only in the mid-2000s (Dwork, 2006; Dwork et al., 2006). The widespread availability of auxiliary information in the Internet age, together with the availability of vast computational power even in a single laptop, completely transformed the landscape of security and privacy, giving substance to what had previously been considered merely theoretical threats.

In summary, if one pictures a system for statistical analysis of confidential data as having the data and all computing devices operating on these data in a vault, with only the results of the analyses emanating from the vault, then security threats are the moral equivalent to breaking into the vault. These threats range from the physical security of the data—from fire, floods, theft of storage devices—through more esoteric information channels, such as analysis of very fine-grained time and power analysis consumption profiles of computations and data accesses. Security threats can often be addressed by appropriate use of back-ups and cryptographic techniques.

Inferential Disclosure: Threats to Individual Privacy

Attacks on individual privacy use legitimately obtained statistical findings, aggregations, and conclusions drawn from confidential data to obtain information about an individual or organization. As in Chapter 4, we focus here only on individual privacy breaches to the data providers—that is, breaches that could not have occurred had the individual's or organization's data not been in the confidential dataset. Inferences based on facts about the population as a whole are not viewed as individual privacy breaches.

Different adversarial goals may require different resources, so to fully specify an attack, one also has to specify the resources to which the adversary has access. Examples of resources include computational capabilities and additional, or auxiliary, information that cannot be obtained by interacting with the dataset. Examples of useful auxiliary information might be personal details about an individual known, for example, to a sibling or coworker, such as the approximate dates on which one has watched a few movies on Netflix (Narayanan and Shmatikov, 2008) and blog entries describing a favorite recent purchase (Calandrino et al., 2011).

Re-identification and De-anonymization

The technical literature and popular press frequently refer to re-identifying data. Such references refer to an approach to privacy protection in which individual data records, containing explicit identifying information, have presumably been de-identified or anonymized, and re-identification refers to reversing this step, tracing an individual record back to its human source. Thus, in a re-identification attack, the goal of the adversary is to trace an individual record to its source. While re-identification may seem difficult for those who do not have access to the underlying data, that is not the case: anyone looking at supposedly de-identified data who is also in possession of auxiliary information about a member of the dataset may well be in a position to re-identify. One example is linking public records, such as voter registration rolls, with de-identified data (Sweeney, 1997; Narayanan and Shmatikov, 2008). Indeed, the richer the dataset, the greater the set of possibilities for useful auxiliary information, and a host of results suggest that de-identified data are either not de-identified or no longer can serve as data. In the words of the President's Council of Advisors on Science and Technology (2014, p. 38):

> Anonymization of a data record might seem easy to implement. Unfortunately, it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data.

We now turn to privacy attacks that can be launched against collections of statistics. We formalize the computational model for interacting with a dataset. In this model, a data analyst can be viewed as the adversary, for several reasons. First, this approach fosters the democratization of statistical research, in theory permitting anyone—a journalist, a concerned citizen, a lawyer in a class action lawsuit—to analyze data. Thus, the computational model does not need to understand the motives of those accessing the data and how their different goals might interact. Second, it recognizes that even the best intentioned data analysts may fail to understand the privacy risks inherent in their own statistical analyses. Third, the approach captures the fact that the choice of measurements selected that is based on exploratory data can result in a privacy breach.[4] For example, suppose the dataset contains detailed demographic information about individuals and their charitable contributions. Suppose further that a member of a specific demographic group is an extremely generous outlier. The analyst, intrigued by the data of the outlier, may choose to examine the correlation between membership in this specific demographic group and charitable donations. Thus, the choice of measurement (correlation between group membership and charitable donation) depends on the data of a specific individual in the dataset, which has the potential to result in an individual privacy breach when combined with other sources of information. Finally, data analysis results in observable actions, such as the publication of statistics and technical papers describing the findings. Taken together, statistics and findings obtained by multiple perfectly trustworthy data analysts can, in combination, compromise privacy.

The problem, then, is how to permit analysts to carry out their work while at the same time understanding that they may be (intentionally or otherwise) a privacy adversary. One solution is for raw data to be hidden from the data analyst, whose access is restricted to repeatedly posing a "query" and receiving a possibly noisy response. Here the word "query" means a function, an algorithm, or a statistical estimation that takes data as input and produces a numerical or categorical output. The noise comes from randomness introduced by the query-answering mechanism, which is introduced in order to protect privacy. Sometimes, the data themselves are altered prior to analysis. This is a generalization of the well-known randomized response technique developed in the 1960s to facilitate surveying of embarrassing or even illegal behavior (Warner, 1965; Erlingsson et al., 2014; Apple, 2016). In this approach, the query-answering mechanism receives data for which privacy has already been "rolled in" through the

[4]Exploratory data analysis, in which the questions asked depend on the data themselves, has other problems, such as overfitting a model to the dataset, meaning that the model does not generalize to the population from which the data were drawn.

randomization process. Often, the data remain unaffected prior to the analysis: that is, the query-answering mechanism has access to the raw, unadulterated data. The queries may be specified ahead of time, for example, when a government agency decides on a set of tables to release and the statistical estimates themselves have added noise. Alternatively, given a specified set of quantities of the dataset to be revealed—such as means and variances—while preserving privacy, synthetic data may be generated whose approximate means and variances (and other prespecified quantities) closely match those of the original dataset. That is, a synthetic dataset can be constructed from existing records on the basis of statistical models that induce noise in statistics relative to those from the original data.

We next turn to queries of the form, "What fraction of the records in the dataset satisfy property *P*?" These are called fractional counting queries or statistical queries. For example, in a database containing information about people's height, weight, and age, a statistical query might ask, "What fraction of the members of the dataset are over 6 feet tall?" The mechanism might compute the true answer to the query and produce as output the sum of the true answer and some random noise. Thus, the mechanisms need only to provide approximate answers to these queries. This approximation is a source of "inaccuracy." Of course, there are many sources of inaccuracy in statistics based on sample surveys, such as sampling error. Probability sampling techniques provide the ability to measure the nature and extent of uncertainty in inferences to the full population. So too with the privacy-enhancing statistical techniques described here. When the inaccuracies are systematic and well behaved, they can be addressed with statistical techniques. This is an important point: many privacy-enhancing techniques introduce errors to protect individual privacy. When this is done using ad hoc and secret techniques, the data analyst cannot compensate and cannot, for example, form correct confidence intervals. In contrast, when the randomization is done using publicly known techniques, a statistician can understand the extent of uncertainties in estimates from the data.

Deliberately introducing inaccuracies is necessary because failure to do so while providing estimates from statistical analyses results in vulnerability to several kinds of attacks on individual privacy. We next describe two of these: reconstruction and tracing attacks.

Reconstruction Attacks

Suppose each data record contains a good deal of nonprivate identifying information and a secret binary digit (bit), one per individual, for example, indicating whether or not the individual has one of the genes associated with Alzheimer's disease. The goal in a reconstruction attack is to determine the secret bits for nearly everyone in the dataset. Reconstruc-

tion attacks succeed, for example, if an adversary can reconstruct the secret bits for all but 1 percent of the records in the dataset.

Reconstruction attacks can be launched against a subset of the rows of a dataset, for example, on the members of an extended family (e.g., parents, children, grandparents, cousins) by proper formulation of the query, such as: "Among the rows in the dataset corresponding to members of the extended family F, what fraction correspond to people over 6 feet tall?"

There is now conclusive research showing that any mechanism providing overly accurate answers to too many counting queries (like the computation of related percentages) succumbs to a reconstruction attack. This collection of results is known informally as the fundamental law of information recovery (or reconstruction). In fact, there is a single attack strategy that succeeds against all such overly accurate answering of too many queries. Here, "too many" is quite small: only n queries are required, where n is the size of the set under attack (which can be as large as the whole dataset or as small as the extended family F). "Overly accurate" means having error on the order of the square root of N, the size of the set under attack.[5] In fact, when n is very small, it is possible to launch a simple attack requiring 2^n queries; in this case, the attack works even if the noise is on the order of n itself.[6] For example, 11 billion estimates are produced from the American Community Survey; it is worth considering the possibility that these estimates can be used to carry out a reconstruction attack on some portion of the respondents.

Tracing Attacks

Reconstruction represents success on the part of the adversary, or, conversely, failure of a privacy mechanism. Tracing—that is, determining whether or not a specific individual is a member of a given dataset—is a much more modest adversarial goal: there are settings in which tracing attacks are possible, but reconstruction attacks are provably impossible. Tracing entered the public consciousness when a group of genomics researchers showed how to use allele frequency statistics in a genome-wide association study, together with the DNA of a target individual and allele frequency statistics for the general population (or a control group), to determine the target's presence or absence in the study (Homer et al., 2008). In response, the National Institutes of Health and the Wellcome Trust changed the access policy to statistics of this type in the studies they

[5]One can think of this as errors on the scale of the sampling error in an opinion poll. An attack requires polling large and overlapping subgroups of the dataset and combining the results in clever fashion.

[6]For a survey of reconstruction results, see Dwork et al. (2017).

fund (see Dwork et al., 2017, for a survey of tracing results; also see further discussion below).

INFERENCE CONTROL TECHNIQUES

Previous National Research Council studies (2005, 2007) have examined techniques and approaches agencies have used for protecting the confidentiality of data while providing access to researchers, and these are briefly described in the panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b). In this section, we elaborate on that discussion with additional perspectives and frameworks for protecting data as a foundation for our discussion of implications and next steps in the final section of this chapter. Specifically, we review the major approaches that statistical agencies have used, including applying statistical disclosure limitation methods and perturbing the data and creating synthetic datasets for analysis, as well as noting the weaknesses of these approaches. We also review different approaches that limit access to data depending on its sensitivity and that require researchers to go to data enclaves in order to analyze the data. We conclude the section with a discussion and review of differential privacy and its potential to meet the needs of statistical agencies.

Statistical agencies and their contractors currently use a variety of statistical disclosure control techniques to attempt to protect the confidentiality of respondents' information (for a summary, see Federal Committee on Statistical Methodology, 2006). Many of them are combinations of techniques. Typical methods used in statistical disclosure control (see Hunderpool et al., 2012; Karr and Reiter, 2014; Skinner, 2009) include

- only releasing information at high levels of aggregation;
- top- or bottom-coding data, by reporting an income of say, $150,000 for all respondents with incomes higher than that figure;
- swapping data, in which variables from pairs of different records are interchanged;
- adding noise to individuals' responses or to computed statistics; and
- creating synthetic data, in which models fit to the real data are used to generate artificial data that are then released.

Many statistical agencies or their contractors also have disclosure review boards, which review data products—including tables, reports, and microdata—prior to release to ensure that the releases do not reveal any information about respondents. The Census Bureau provides instructions to

program managers and a checklist to complete and submit to its disclosure review board.[7]

De-identification Through Perturbing the Original Data

"De-identification" typically refers to a situation in which there is a collection of records, each essentially belonging to or being about a specific individual. In this technique, certain data fields in an individual record are eliminated, coarsened, or otherwise modified, leaving the remaining fields untouched (or perhaps more lightly modified). De-identification does not prevent linkage attacks, in which a second database is used as auxiliary information to compromise privacy in the first database. This type of attack is at the heart of the vast literature on hiding small cell counts in tabular data (cell suppression). Naïve de-identification resulted in perhaps the world's most famous linkage attack, in which the medical records of the governor of Massachusetts were re-identified by linking voter registration records to anonymized Massachusetts Group Insurance Commission medical encounter data, which retained the date of birth, gender, and ZIP code of the patient (Sweeney, 1997). Sweeney proposed an antidote, known as *k*-anonymity. In *k*-anonymity, a syntactic condition requires that every "quasi-identifier" (essentially, a combination of nonsensitive attributes) must appear at least *k* times in the published database if it occurs at all. This can be achieved by coarsening attribute categories, for example, replacing 5-digit ZIP codes with their 3-digit prefixes. In addition to potentially being hard to compute (in the sense discussed above), the choice of category coarsenings may reveal information about the database. That is, had the data been different, the coarsenings chosen would have been different, thus the choice itself may constitute an individual privacy breach. Perhaps more concerning is the fact that if a given individual is known to be in the dataset, then *k*-anonymity permits an attacker to narrow down to a set of the size of no more than *k* the set of records possibly corresponding to the individual. If, for example, all *k* of these records share a diagnosis, then the individual necessarily received this diagnosis. Difficulties of this sort led to a series of works and additional syntactic constraints on anonymized datasets (Machanavajjhala et al. [2006] on *l*-diversity; Xiao and Tao [2007] on *m*-invariance; Li et al. [2007] on *t*-closeness).

In general, syntactic conditions may fail to capture the semantics of privacy: that is, suppressing and aggregating various data fields is not formally tied with a mathematical definition of privacy. The methodological problem with syntactic considerations is understanding when these syntactic condi-

[7]See https://www.census.gov/srd/sdc/wendy.drb.faq.pdf and https://www.census.gov/srd/sdc/drbchecklist51313.docx [August 2017].

tions protect against individual privacy breaches. It is also important to note that the damage done to the data in the anonymization process can completely destroy utility (Brickell and Shmatikov, 2008).

Synthetic Data

Using synthetic data is an approach to privacy protection in which a model is first specified and estimated from the data and then fictitious samples are generated according to the model (Reiter and Raghunathan, 2007; Rubin, 1993).[8] For example, if one assumes the data are drawn from a normal distribution, one learns from the data a mean μ and a variance σ^2 and then generates random samples drawn from the normal distribution $N(\mu,\sigma^2)$. These random samples are assembled into a dataset that is publicly released. The Census Bureau has released synthetic datasets for the Survey of Income and Program Participation linked to Social Security income records (Benedetto et al., 2013) and the Longitudinal Business Database (Kinney et al., 2011).

Although at first glance this approach appears to protect privacy—after all, only random noise is released—the parameters μ and σ^2 themselves leak information about the original dataset. Synthetic data do not protect privacy merely by virtue of being synthetic; the process by which they are generated must also be protective.[9] Another limitation is that synthetic data can tell no more about a dataset than can be encapsulated by the model. The model provides assurance that a class of models, which are distinct from those used in generating the synthetic datasets, can be estimated with expected statistical error properties. Furthermore, in order to obtain estimates of the added uncertainty of statistical values arising from the synthetic generation of data, multiple synthetic datasets are often generated, complicating the statistical estimation and, potentially, increasing risk of privacy loss.

Weaknesses of Statistical Disclosure Limitation Methods

Although the methods discussed may protect the data against some attacks, they cannot be guaranteed to protect the privacy of respondents, and they all negatively affect the utility of the data for researchers (see, e.g., Reiter, 2012). As argued above, aggregated tables can be assembled

[8]Synthetic data is a general term. The model-based, multiple imputation approach of Rubin (1993) and others is one way to create synthetic data; other approaches can be found in the differential privacy literature.

[9]See, for example, https://obssr.od.nih.gov/synthetic-data-protecting-data-privacy-in-an-era-of-big-data/ [August 2017].

to provide information about ever-smaller segments of the population. As noted above, 11 billion estimates are published every year from the American Community Survey, with the data sliced and diced on multiple dimensions for many different geographic areas (see Hedrick and Weister, 2016). An adversary can assemble different statistics to learn about individuals in the dataset. Data swapping can distort multivariate relationships in the data and result in unusual patterns, for example, if the income of a fast food worker is swapped with that of a neurosurgeon (although in practice restrictions would be placed on preserving certain univariate and multivariate relationships while swapping). Synthetic data convey only the information in the model used to generate them, and that information can be disclosive. In addition, synthetic data cannot be used to discover any information not in the model, and it might therefore be questioned why one would not just publish the model.

Approaches for Tiered Access

In addition to the methods described above, the other major approach statistical agencies use to protect the privacy of data is to restrict access to and use of the data with legally binding contracts and administrative procedures. Given the different levels of legal protection for different data (see Chapter 4), there have also been efforts to tailor the protections for a particular dataset to the sensitivity of the information and the potential harms that could result from disclosure. Sweeney and her colleagues (2015) have "tagged" data with different grades of security measures needed to access the data (see Box 5-1). By using a tiered access system, data are able to be matched with appropriate privacy control techniques. By using tiered access systems, organizations are able to grant different levels of access based upon the merit of individuals using the data for research.

Altman and colleagues (2015) elaborate on several privacy controls, including procedural, technical, educational, economic, and legal means, and they note how these can be applied throughout the information life cycle of collection/acceptance, transformation, retention, access and release, and post-access (see Box 5-2 and Table 5-1). Their approach to tiered access is to evaluate identifiability against potential harms from uncontrolled use of the data (see Figure 5-1). In the lowest risk category, data that are not harmful or could result in minor harm would be available for public use. The next tier, either data that contain identifiers but are not harmful or data that are de-identified that could result in significant harm, would be de-identified and require additional steps, such as consent and terms of service. The final two tiers cover data that could result in major harms and would work in similar ways, involving formal application and data use

BOX 5-1
Datatagging: Example

The datatag system (Sweeney et al., 2015) was created to maximize data sharing while minimizing risk of harm from sensitive information. Through the notion of datatags, data are allowed to be accessed and handled with different security and credential measures. The system contains six different grades of security described by numbers or colors, from one (blue) to six (crimson), where one contains the least sensitive information.

- A blue datatag (one) requires no access credentials to access. The data typically do not contain any personal information. Examples include non-personal research data that has already been published and public-use files.
- A green datatag (two) is for controlled public data. These data can be accessed by the public with minimal agreement, but access often requires verification, similar to providing a valid email address.
- A yellow datatag (three) is the first level for which a data use agreement is required. This agreement often is in click-through form, and a password and approval are required in order to access data. This is also the first stage at which data are encrypted when transmitted.
- An orange datatag (four) is the first level for which a written signed agreement is required for data access. At this stage, the data are also encrypted in storage.
- A red datatag (five) usually requires two-factor authentication for access, such as a phone number and email address.
- A crimson datatag (six) is similar to a red datatag, except data are multi-encrypted in storage.

There are multiple methods to determine at what level data should be tagged. One possibility includes having an interview with a panel, such as an Institutional Review Board, which includes privacy experts. It is also possible to have automated data tagging by computer programs. It is recommended to have some sort of human-assisted approaches in combination with computer techniques, however, to tag data. Sweeney and colleagues (2015) draw a comparison with TurboTax, in which a user goes through a series of "yes" or "no" questions for most determinations while experts are available to address questions.

BOX 5-2
Privacy Control Techniques: Example

In trying to create a new privacy framework for government data, Altman and colleagues (2015) delineated five different privacy controls that are available to policy makers: procedural, technical, educational, economic, and legal. These privacy controls can also be used in conjunction with five stages of data use: collection and acceptance, transformation, retention, access and release, and post-access (see Table 5-1 for more information).

1. Procedural controls involve internal procedures, such as implementation notices, creating inventories, or vetting internal or external access to an organization's database. Across data life cycles, examples of procedural means include data minimization and access controls.
2. Technical controls are defined as statistical methods, computational methods, and human factors analysis. Examples include synthetic data, encryption, and differential privacy.
3. Educational controls include providing information to data subjects, collectors, controllers, and recipients about privacy practices and risks. Techniques that can be used for education include notices and transparency.
4. Economic controls include any intervention intended to change the economic incentives of stakeholders, such as fees to access data, provision of insurance, and property rights agreements.
5. Legal controls ensure data are properly protected through legal rights among stakeholders. Legal means can include data rights, terms of service, and civil or criminal punishment.

agreements. The most sensitive data would be held in enclaves and require strict data logs, as major harms could result if the person was identified.

Data Enclaves

Federal statistical agencies provide access to some microdata files through the Federal Statistical Research Data Centers (FSRDCs).[10] In these controlled environments, researchers can conduct statistical analysis on specific datasets for their approved projects. Prior to being permitted to use an FSRDC, researchers are required to go through screening and training on preserving the confidentiality of the data and are sworn to protect the data, with criminal penalties for disclosure. Researchers seeking to link external data to data in the FSRDC must provide the data to the Census Bureau,

[10]See https://www.census.gov/fsrdc [August 2017].

which does the linkage. No data are brought into or out of the FSRDC itself because the hardware used is a thin client, which is built for remote access to a server, with a virtual private network. Researchers can use a variety of statistical software, but procedures that would permit viewing individual records are disabled. No results are permitted to be taken outside of an FSRDC until they have gone through disclosure review. Similar procedures are used at other data enclaves (see Box 5-3). There have been no known security breaches of these secure enclaves; however, the limited access can make it difficult to replicate the research from their use (Altman et al., 2015).

Differential Privacy

Differential privacy is a definition of privacy tailored to the statistical analysis of large datasets. Differential privacy precisely captures the difference between learning about an individual in the dataset and learning about a population (as represented by a dataset) as a whole. Differentially private algorithms ensure that the only possible harms are group harms: the outcome of any analysis is almost equally likely independent of whether any individual joins, or refrains from joining, the dataset. If the same output occurs whether or not a particular person is in the dataset, then that person cannot suffer any privacy loss.

This seems paradoxical. If the behavior of the algorithm is the same even when one changes the data on which it is operating, how can one understand the results? The answer to the seeming paradox is through the careful introduction of randomness into the computation. For example, if 52 percent of the population supports a candidate and one randomly samples a large number of people in the population, one can expect that the fraction of the sample supporting the candidate will be close to 52 percent. The sampling process introduces randomness in the answer, but the answer is still very likely to be roughly the same, as long as the privacy loss parameter is relatively small.

Differential privacy is a property of an algorithm, not an output of the algorithm. Differentially private algorithms are randomized, meaning that they intentionally introduce randomness into the computation and therefore the statistical estimates produced. Randomness is used in essentially all cryptographic algorithms: for example, randomness is used in choosing secret keys for digitally signing documents; the fact that they are generated randomly for each signer is what makes them secret, preventing forgeries.

In looking at differential privacy, one also needs the concept of adjacent datasets: x and y are adjacent datasets if one is contained in the other and the larger one contains exactly one additional record. For example, the records in the dataset might be individual hospitalization histories: the

TABLE 5-1 Privacy Controls over Data Life Cycle

Life Cycle Stage	Procedural	Economic	Educational	Legal	Technical
Collection/ Acceptance	Collection limitation; data minimization; data protection officer; Institutional Review Boards; notice and consent procedures; purpose specification; privacy impact assessments	Collection fees; markets for personal data; property right assignment	Consent education; transparency; notice; nutrition labels; public education; privacy icons	Data minimization; notice and consent; purpose specification	Computable policy
Transformation	Process for correction		Metadata; transparency	Right to correct or amend; safe harbor de-identification standards	Aggregate statistics; computable policy; contingency tables; data visualizations; differentially private data summaries; redaction; statistical disclosure limitation techniques; synthetic data
Retention	Audits; controlled backups; purpose specification; security assessments; tethering		Data asset registries; notice; transparency	Breach reporting requirements; data retention and destruction requirements; integrity and accuracy requirements	Computable policy; encryption; key management (and secret sharing); federated databases; personal data stories

Access/Release	Access controls; consent; expert panels; individual privacy settings; presumption of openness vs. privacy; purpose specification; registration; restrictions on use by data controller; risk assessment	Access use fees (for data controller or subjects); property rights assignment	Data asset registries; notice; transparency	Integrity and accuracy requirements; data use agreements (contract with data recipient, terms of service)	Authentication; computable policy; differential privacy; encryption (including functional and homomorphic); interactive query systems; secure multiparty computation
Post-Access (audit, review)	Audit procedures; ethical codes; tethering	Fines	Privacy dashboard; transparency	Civil and criminal penalties; data use agreements (terms of service); private right of action	Computable policy; immutable audit logs; personal data stores

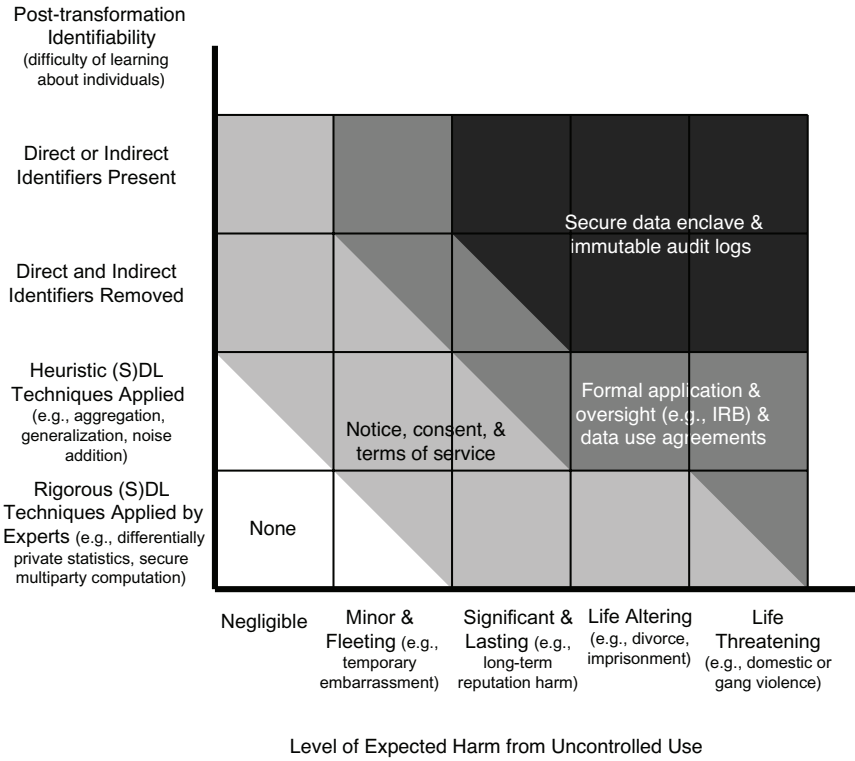

FIGURE 5-1 Privacy controls needed given the identifiability of the data and the expected harm from uncontrolled use.
SOURCE: Adapted from Altman et al. (2015, p. 2046).

larger dataset will contain the hospitalization history of exactly one additional person. Differentially private algorithms ensure that the probability of any event is essentially equally likely when the algorithm is running on adjacent databases. "Essentially equally likely" is measured by a privacy loss parameter, usually denoted as epsilon, ε. Smaller values of ε give better privacy protection than larger values.

Think of ε as a small number, say, 0.01. Let U denote the universe of theoretically possible data records. Formally, letting $M(x)$ denote the result when algorithm M is run on database x, one can define differential privacy as follows (Dwork et al., 2006): a randomized algorithm $M{:}U^* \rightarrow \text{Range}(M)$ is ε-differentially private if for all adjacent datasets x,y and all events $S \subseteq \text{Range}(M)$, we have $\Pr[M(x) \in S] \leq e^{\varepsilon} \Pr[M(y) \in S]$. When ε is small, $e^{\varepsilon} \approx 1+ \varepsilon$. In this case, $\Pr[M(x) \in S] \leq (1 + \varepsilon) \Pr[M(y) \in S]$, meaning that the

BOX 5-3
The Five Safes Framework: Example

The five safes framework was developed by the data service in the United Kingdom and is used in many research labs, including the country's Office for National Statistics and the Administrative Data Research Network (Administrative Data Research Network, 2017). The purposes of the five safes are to have protections at different levels to minimize the risk of sensitive and confidential data being used incorrectly. The five safes are safe projects, safe people, safe settings, safe data, and safe outputs.

Safe projects: Researchers submit proposals that must be approved in order to begin research. A proposal must meet a public good criterion and must hold a certain amount of value in order to receive acceptance.

Safe people: Once a project is approved, researchers must explain why they are suitable for data access, must be affiliated with a respected academic institution, and must sign a user agreement for data use. They must complete a mandatory training course about the legal and ethical use of confidential and sensitive data, statistical disclosure control, and using the secure labs.

Safe settings: After training, researchers can access data only through safe settings, such as remote access through Citrix secure remote access technology, technology that is currently used by both military and banking sectors. Ethical hackers are employed in order to test the security of the servers, in order to ensure that unauthorized access to data is as limited as possible.

Safe data: Safe data attempts to limit disclosure risk of the data as much as possible. Due to all previous precautions, data at this stage are the most vulnerable for re-identification. Therefore, no data are able to be taken out of the safe settings.

Safe outputs: After using data, researchers cannot take any results out of the safe settings without statistical disclosure techniques being applied to the output to reduce the risk of re-identification.

All of the five safes aim to maximize researcher access to important data while preventing certain forms of individual privacy breaches or threats.

two probabilities are nearly equal. The definition says that the probability of M observing any output is nearly equal, independent of whether the database is x or y.

For example, suppose x is a database of medical records and y is the database consisting of x and the data of one additional person, "George." Assume that M is an algorithm for computing correlations between the pollutant in a small town and chronic bronchitis. Now, George is in y but not in x. Thus, the output $M(x)$ is not "about" George since his data were not inputs to M. Note, however, that since M might be about people in general, it can have implications for George. For example, it could be used to inform

a doctor of basic scientific information that will be useful in diagnosing George's respiratory difficulties. Differential privacy ensures that when the algorithm M is run on database y (which does contain George's medical record), essentially nothing more is learned about George than was learned when M is run on database x (which does not contain George).

As a further illustration, let M be a 0.1-differentially private mechanism (i.e., an ε-differentially private mechanism where ε has value 0.1) reporting the number of people in the database suffering from chronic bronchitis. Suppose we tell a gambler all of database x as well as the entirety of George's medical record, and suppose further that George suffers from chronic bronchitis. This means that the true answer to the question, "How many people in the database suffer from chronic bronchitis?" is larger—by exactly 1—when the database is y. We then flip an unbiased coin and if it comes up heads, we run M on input x, while if it comes up tails, we run M on input y. Suppose M outputs the number 501. If we do not give the outcome to the gambler, he has a probability of exactly 0.5 of guessing which input—x or y—we used. If we tell him the output his probability of correctly guessing which input we used increases from 0.5 to at most 0.525. By hiding the data of each individual in this way, differentially private algorithms ensure that anything that is learned is about people in general and not about individuals in the dataset. Differentially private reporting of allele protein frequencies will prevent even someone who has obtained a sample of George's DNA from determining whether or not George is in the case group.

Weaknesses of Differential Privacy

Differential privacy is a mathematically rigorous concept, so algorithms can be designed and proven to provide it. The approach automatically provides protection against linkage attacks, as well as present and future auxiliary information. It also provides a formal measure of privacy loss and a calculus for computing how privacy losses compound over multiple data analyses. In addition, it yields a collection of simple privacy-preserving building blocks that can be combined to yield differentially private versions of sophisticated analytical techniques from statistics and machine learning.

Nonetheless, differential privacy is not a panacea in many ways. First, neither differential privacy nor any other technique can circumvent the fundamental law of information recovery: overly accurate estimates of too many statistics can completely destroy privacy. Thus, differential privacy is not an immediate solution to the problem uncovered in the genome-wide association study discussed above; it must be combined with sophisticated techniques for false discovery rate control so as to limit the number of (noisy) statistics actually revealed. Again, there are limits for *any*

technique that provides any reasonable notion of accuracy (Dinur and Nissim, 2003; Dwork et al., 2007, 2015; Muthukrishnan and Nikolov, 2012; Kasiviswanathan et al., 2013). Differential privacy often achieves the best possible bounds, and it does provide provable guarantees. Unlike other techniques, differential privacy provides the tools needed to measure and bound the cumulative risk as the data are used and reused, permitting the informed enforcement of a privacy "budget." In other words, all systems are eventually in trouble: differential privacy allows one to monitor the trouble and decide when to stop.

Conceptually, there is no difference between multiple releases of synthetic datasets and interactive query systems (sometimes the query is simply "give me a synthetic dataset capturing the following attributes"). Nonetheless, the engineering and social challenges should not be underestimated. To adhere to a privacy budget, it will be necessary to optimize the choice of queries to be handled and to determine how to use the privacy budget as effectively as possible while ensuring fairness of access to many users of the data, not all of whom are known in advance. For example, one needs to protect against attacks intended to deplete the privacy budget just to shut down access to an agency's data. One also needs to remove side-channel attacks and other threats. All of these challenges exist for the other approaches to protecting privacy; however, there are no tools to measure them.

Second, in traditional data analysis an analyst looks at the dataset. Differential privacy limits interaction with the dataset, not allowing the analyst to see the raw data (although she may have auxiliary information about the dataset, as in our example above about George and the sets x and y). Moreover, data analysts are not trained to operate in this scenario. They do not know about cumulative privacy loss, nor are they experienced in considering a computation to minimize privacy loss while maximizing utility.

Differential privacy also hides outliers. Indeed, it is often the outliers who most need protection. Note that many kinds of analyses make use of outliers (e.g., income distributional statistics).

Differential privacy may require larger sample sizes in order to learn things about the full population that, without privacy concerns, would require fewer sample cases.

Differential privacy intentionally introduces statistical noise. This is essential, but using currently known techniques the noise may be too much to permit useful learning unless the dataset is very large. In some kinds of analyses, differential privacy introduces exactly the minimum amount of noise needed to prevent reconstruction and tracing attacks; this means no technique exists for adding less noise in these cases. However, not all sets of analyses are so well understood. In these cases, the situation may be better than assumed: not only is it possible that better differentially private algo-

rithms exist than those currently known, but there might even be a different and meaningful definition of privacy that can achieve better accuracy than anything known. It is also possible that the situation is worse than assumed: it is possible that any algorithm that adds less noise than the existing ones could be shown to be vulnerable to some new method of statistical attack.

Differential privacy algorithms come equipped with a privacy parameter (ε). The meaning of this parameter is not well understood. The choice of the parameter is currently not assigned to any actor in the federal statistical system.

Differential privacy is a young field, and good algorithms—with high accuracy and low privacy loss—for many key analytical tasks are still under development. The federal statistical system produces statistics that are simple means and totals, such as sums of weighted products of variables, but also correlations, regression model coefficients, and cumulative statistics on multivariate distributions. Much remains to be developed to apply differential privacy techniques to some of these statistics.

Differential privacy does not distinguish between the types of information protected: hair color implicitly receives the same treatment as sexual identity. The privacy guarantee is framed in terms of a comparison between the behavior of the algorithm when one person (with any color hair and any sexual preference) is in the dataset compared with the behavior when this same person is absent. This may seem silly: Why go to such efforts to protect hair color? But it frees the algorithm designer from the need to know whether or not specific attributes are considered sensitive, either at all or by a specific user. For example, it protects the information about all genetic markers, even those not currently known to be correlated with, say, Alzheimer's disease. If, as has happened in the past, certain genotypes are later discovered to have such a correlation, differential privacy will have protected against that future information.

IMPLICATIONS FOR FEDERAL STATISTICAL AGENCIES

As noted in Chapter 4, federal statistical agencies are required to follow prescriptive guidance from the U.S. Office of Management and Budget and the National Institute of Standards and Technology related to the Federal Information Security Management Act and other requirements for maintaining the security of their information systems. In contrast, agencies draw on general best practices and their staff's professional judgment in protecting their statistical data products from inferential disclosure individual priavacy breaches. Statistical agencies currently use a number of statistical disclosure methods to protect the confidentiality of their data; however, these methods are increasingly susceptible to privacy breaches given the proliferation of external data sources and the availability of high-powered computing that could enable inferences about people or entities in a dataset,

re-identification of specific people or entities, and even reconstruction of the original data. Thus, in our first report we drew the following conclusions:

> **A continuing challenge for federal statistical agencies is to produce data products that safeguard privacy. This challenge is increased by the use of multiple data sources.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Conclusion 5-5, p. 90)

> **As federal statistical agencies move forward with linking multiple datasets, they must simultaneously address quantifying and controlling the risk of privacy loss.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Conclusion 5-6, p. 95)

As noted above and in our first report, differential privacy provides a framework and a measure of privacy loss and tools for tracking this loss over analyses of the data, creating a "privacy loss budget." Federal statistical agencies typically seek to maximize the statistical use of their data within the constraints of ensuring the statistical products do not reveal anything about individual respondents. Agencies have only recently begun to examine the implications of a privacy loss budget for their own production of statistical products and secondary analyses by external users (see, e.g., Abowd et al., 2017). Use of these methods will require new algorithms to be developed, tested, and reviewed not only by the scientific community, but the many stakeholders for federal statistics. Therefore, the panel also made the following recommendations in its first report:

> **Statistical agencies should engage in collaborative research with academia and industry to continuously develop new techniques to address potential breaches of the confidentiality of their data.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Recommendation 5-1, p. 96)

> **Federal statistical agencies should adopt modern database, cryptography, privacy-preserving, and privacy-enhancing technologies.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Recommendation 5-2, p. 96)

However, the panel recognizes that the current era is one of transition: there is awareness of weaknesses of current statistical disclosure limitation methods, but the feasibility for federal statistical agencies of implementing new technologies, such as differential privacy, has not been clearly demonstrated. So far, the only application of differential privacy in federal

statistical products is the Census Bureau's OnTheMap application,[11] which implements a variant of differential privacy allowing users to see where people work and where workers live (see Machanavajjhala et al., 2008). More research and development will be needed to show how this can be done and the costs and benefits of implementation of these methods. It also will require statistical agencies to hire and train staff to use these technologies.

RECOMMENDATION 5-1 Federal statistical agencies should ensure their technical staff receive appropriate training in modern computer science technology including but not limited to database, cryptography, privacy-preserving, and privacy-enhancing technologies.

Overall, much work, interaction, and collaboration will be needed across the various disciplines and stakeholders as agencies seek to move forward to provide stronger privacy protection for the data they either collect from respondents or acquire access to from other administrative and private-sector sources for statistical purposes. It will be critical for there to be robust discussions of the implications of this approach for all stakeholders and these discussions will need to be informed by concrete examples to help everyone understand how use of these technologies will affect them. Pilot studies or test cases will be valuable in identifying the variety of issues that affect agencies and the users of their data, including effects on timeliness of production, the scope of statistical products produced, the utility of the resulting estimates, and the usability of microdata by external researchers.

Comparisons will need to be made using agencies' current procedures with state-of-the-art differentially private algorithms and various levels of epsilon from a variety of federal statistical datasets to evaluate the effects on accuracy of results and utility of the resulting data. Some efforts along these lines have been conducted (see, e.g., McClure and Reiter, 2012), and additional pilot studies using different federal survey datasets would be beneficial and help enhance understanding by the communities involved. More specifically, the panel thinks there is potential value in pilot studies that look retrospectively, as well as prospectively, at the current uses of agency survey datasets and how a privacy budget would have been used over time, across users, and for various activities. For example, agencies could look back at uses over the past 5 years using internal information, as well as literature searches, to assess the volume and types of analyses that have been conducted and how the use of a privacy budget could have affected internal and external analyses of the data or, conversely, the effects on privacy loss given the uses.

The panel also thinks that much could be learned by generating syn-

[11] See https://onthemap.ces.census.gov/ [August 2017].

thetic data in a differentially private manner and encouraging researchers and students to conduct their analyses on this dataset and use verification servers (see Karr and Reiter, 2014) to assess the fidelity of the results with the original dataset. Note, however, that even synthetic data are subject to the fundamental law of information recovery: if a synthetic dataset permits overly accurate estimates of too many statistics, then privacy could be destroyed. For this reason, query-response systems may deliver more accuracy for queries actually of interest than what can be achieved using synthetic datasets, at the same level (epsilon) of privacy loss.

The panel suggests the use of an epsilon registry to encourage those who work with or profit from the use of personal information to take a greater interest in the algorithms and share their experience in setting this parameter so that this work moves beyond the community of privacy researchers (Dwork and Mulligan, 2014). A typical element in the registry might describe a use of differential privacy, including the value of ε used (this could be done in several ways, such as per calculation, together with the number of analyses run, or a "burn rate" of privacy loss per day or week); a discussion of the factors leading to this choice; the granularity at which differential privacy is applied (per data attribute or per individual record, which typically contains many attributes); whether or not data are "retired" when a privacy loss limit is reached; and the variant of differential privacy used.

Ultimately, the adoption of new technologies will be driven either by necessity or the broad embrace of the communities of researchers, data users, and privacy advocates. The panel hopes that agencies will engage in collaborative research programs with the academic and user communities to promote greater use and understanding of new and emerging privacy-protection methods. We also hope that workshops and other public discussions will be held on the results of the research and the policy issues associated with setting and allocating privacy budgets. These issues go beyond the purview of a single statistical agency and could have dramatic effects on the uses, users, and providers of statistical data. The implications of these issues need to be broadly discussed and a transparent and participatory process outlined for moving forward.