Differential Privacy: A Primer for a Non-technical Audience*

(Preliminary version)

Kobbi Nissim†[1], Thomas Steinke[2], Alexandra Wood[3], Micah Altman[5], Aaron Bembenek[6], Mark Bun[2], Marco Gaboardi[4], David R. O'Brien[3], and Salil Vadhan[2]

[1]Department of Computer Science, Georgetown University.

[2]Center for Research on Computation and Society, Harvard University.

[3]Berkman Klein Center for Internet & Society, Harvard University.

[4]State University of New York at Buffalo.

[5]Program on Information Science. Massachusetts Institute of Technology.

[6]School of Engineering and Applied Sciences. Harvard University.

May 7, 2017

Keywords: differential privacy, data privacy, social science research

*This document is the product of a working group of the *Privacy Tools for Sharing Research Data* project at Harvard University (http://privacytools.seas.harvard.edu). The working group discussions were led by Kobbi Nissim. Kobbi Nissim, Thomas Steinke, and Alexandra Wood were the lead authors of this document. Working group members Micah Altman, Aaron Bembenek, Mark Bun, David R. O'Brien, Marco Gaboardi, Kobbi Nissim, Thomas Steinke, Salil Vadhan, and Alexandra Wood contributed to the conception of the document and to the writing. We thank Caper Gooden, Deborah Hurley, Georgios Kellaris, Daniel Muise, and Michel Reymond for their many valuable comments on earlier versions of this document. This material is based upon work supported by the National Science Foundation under Grant No. 1237235, as well as by the Alfred P. Sloan Foundation.

†Work towards this document was completed while the author was at the Center for Research on Computation and Society at Harvard University.

Abstract

Differential privacy is a formal mathematical formal mathematical framework for guaranteeing privacy protection when analyzing or releasing statistical data. Recently emerging from the theoretical computer science literature, differential privacy is now in initial stages of implementation and use in various academic, industry, and government settings.

This document is a primer on differential privacy. Using intuitive illustrations and limited mathematical formalism, this primer provides an introduction to differential privacy for non-technical practitioners, who are increasingly tasked with making decisions with respect to differential privacy as it grows more widespread in use. In particular, the examples in this document illustrate ways in which social science and legal audiences can conceptualize the guarantees provided by differential privacy with respect to the decisions they make when managing personal data about research subjects and informing them about the privacy protection they will be afforded.

Contents

Audience

The goal of this primer is to introduce the reader to the concept of *differential privacy*, a new formal mathematical model of privacy protection. Differential privacy is used in some of the privacy-preserving tools for social scientists being developed by the *Privacy Tools for Sharing Research Data* project at Harvard University,[1] as well as many other projects across academia and industry, including implementations by statistical agencies such as the U.S. Census Bureau and companies such as Google and Apple.[2]

This document is written for a broad, non-technical audience, with the goal of presenting a generally accessible overview of the mathematical concepts underlying differential privacy. While this article is written with *analysts* of privacy-sensitive data in mind, some sections take the point of view of a *data subject*, i.e., an individual whose personal data are used in a statistical analysis. The perspective of the data subject is used, in particular, where we discuss how differential privacy controls the increase in risk to individuals due to the contribution of their privacy-sensitive data to a data analysis.

We hope that this way of describing the features of differential privacy will help social science researchers understand the guarantees provided by differential privacy, informing future decisions regarding whether to use differential privacy in their research process and, if so, what types of promises they should make to their research subjects about the guarantees differential privacy provides. In addition, these illustrations are intended to help legal scholars and policymakers consider how current and future legal frameworks and instruments will apply to tools based on formal privacy models such as differential privacy. For step-by-step guidance on using differential privacy, additional resources, such as documentation for individual software implementations relying on differential privacy, should be consulted.

1 Introduction

A common challenge in empirical social science is the sharing of privacy-sensitive data for the purposes of replication and secondary research. Social science research data often contain personal information about individual participants that is considered sensitive or confidential. Improper disclosure of such data can have adverse consequences for a research subject's relationships, reputation, employability, insurability, or financial status, or even lead to civil liability, criminal penalties, or bodily harm. Due to these and related concerns, a large body of laws, regulations, ethical codes, institutional policies, contracts, and best practices has emerged to address potential privacy-related harms resulting from human subjects research.

1.1 Introduction to legal and ethical frameworks for research data privacy

Generally, research policies require researchers to protect privacy as a principle that is fundamental to safeguarding the dignity and welfare of their subjects. Researchers are accordingly responsible for implementing privacy-protective measures and effectively conveying the extent of protection afforded to their subjects. In addition, specific administrative, technical, and physical measures are mandated by privacy laws and the policies of research institutions, funding organizations, and regulatory agencies. Notably for researchers in the United States, research involving human

[1] See Harvard University Privacy Tools Project, http://privacytools.seas.harvard.edu.

[2] See Section 9.2 below for a list of other implementations of differential privacy.

subjects is governed by the Federal Policy for the Protection of Human Subjects, or the Common Rule. When conducting research involving personal information at an institution subject to the Common Rule, a researcher must secure approval from an institutional review board (IRB) and fulfill ethical obligations to the participants, such as disclosing the risks of participation, obtaining their informed consent, and implementing specific measures to protect privacy as required by the IRB.

Additional legal standards for privacy that may apply to research data are found in federal information privacy laws which protect certain categories of information, such as health, education, financial, and government records, among others, as well as state data protection and breach notification laws which prescribe specific data security and breach reporting requirements when managing certain types of personal information. It is also common for universities and other research institutions to adopt policies that require their faculty, staff, and students to abide by certain ethical and professional responsibility standards and set forth enforcement procedures and penalties for mishandling data. Further restrictions apply when privacy-sensitive data are shared under a contract; in fact, the terms of the agreement will often strictly limit how the data can be used or redisclosed by the recipient.

Privacy requirements are also found in technical standards such as those from the International Organization for Standardization, which provides technical guidance on implementing information security controls to protect personally identifiable information. In addition, international privacy guidelines have been adopted by governments across the world. The most widely-followed guidelines are the privacy principles developed by the Organisation for Economic Co-operation and Development, which include collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation, and accountability principles. The right to privacy is also protected by various international treaties and national constitutions.

Taken together, the safeguards required by these legal and ethical frameworks are designed to protect the privacy of research subjects; ensure they fully understand the scope of personal information to be collected and the privacy risks associated with their participation in a study; avoid administrative, civil, and criminal penalties against themselves and their host institutions; and maintain the public's trust and confidence in scientific research.

1.2 Traditional statistical disclosure limitation techniques

A number of technical measures for disclosing data while protecting the privacy of individuals have been produced within the context of these legal and ethical frameworks. A subset of techniques for the release of statistical data have been developed under the title of *statistical disclosure limitation (SDL)* and are widely used by statistical agencies, data analysts, and social science researchers. This term refers to a collection of techniques that are applied to sets of data containing privacy-sensitive personal information with the aim of making it more difficult (or impossible) to learn personal information that is specific to an individual. This category of techniques encompasses a wide range of methods for suppressing, aggregating, and generalizing attributes of individuals in the data. [3] Such techniques are often applied with the explicit goal of *de-identification*, whereby data are transformed by means of redaction or coarsening so as to make it difficult to link an identified person to a record in a data release.

[3]For an overview of traditional SDL techniques, see Federal Committee on Statistical Methodology, Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper 22 (2005), https://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf.

However, changes in the way information is collected and analyzed, including advances in analytical capabilities, increases in computational power, and the expanding availability of personal data from a wide range of sources, are eroding the effectiveness of traditional SDL techniques. Since the 1990s, and with increasing frequency, privacy and security researchers have demonstrated that data that have been de-identified can often be successfully *re*-identified via record linkage [19]. Re-identification via record linkage, or a *linkage attack*, refers to the re-identification of one or more records in a de-identified dataset by uniquely linking a record in a de-identified dataset with identified records in a publicly available dataset, such as a voter registration list. As an example Latanya Sweeney applies such an attack on Group Insurance Commission (GIC) data containing anonymized Massachusetts patient data. Sweeney observed that records in the GIC data contain patients birth date, sex, and zipcode information, and that many of the patients have a unique combination of these three attributes. Re-identification was, hence, possible by joining the GIC data with publicly available Cambridge Voter Registration records which include birth data, sex, and zipcode information alongside Cambridge residents explicit identity. Weaknesses have also been found with respect to other approaches to privacy via the application of a variety of sophisticated attacks. Understanding the limits of these techniques is the subject of ongoing research.

1.3 The emergence of formal privacy models

Re-identification attacks are becoming increasingly sophisticated over time, as are other types of attacks that seek to infer characteristics of individuals based on information about them in the data. Successful attacks on de-identified data have shown that traditional technical measures for privacy protection may, in particular, be vulnerable to attacks devised after a technique's deployment and use. Some de-identification techniques, for example, require the specification of attributes in the data as identifying (e.g., names, dates of birth, or addresses) or non-identifying (e.g., movie ratings or hospital admission dates). They may also require a careful analysis of present and future data sources that could potentially be linked with the de-identified data and enable re-identification of the data. Researchers may later discover that attributes initially believed to be non-identifying can in fact be used to re-identify individuals, or that unanticipated sources of auxiliary information can be used for re-identification. Indeed, the scientific literature provides numerous real-world demonstrations of attacks with results of this nature.

Issues such as these have underscored the need for privacy technologies that are immune not only to linkage attacks, but to any potential attack, *including attacks that are currently unknown or unforeseen.* They have also demonstrated that privacy technologies must provide meaningful privacy protection not only in a "standalone" setting but also in settings in which extensive external information may be available to potential attackers, including employers, insurance companies, relatives, and friends of a subject in the data. In addition, real-world attacks have illustrated that ex-post remedies, such as simply "taking the data back" when a vulnerability is discovered, are ineffective because many copies of a set of data typically exist. As an example, in 2006 AOL published anonymized search history of 650,000 users over a period of three months. Shortly after the release, the New York Times identified a person in the release and AOL removed the data. However, in spite of its removal, mirror copies of the data are accessible on the Internet.

In response to the accumulated evidence of weaknesses with respect to traditional approaches, a new privacy paradigm has emerged from the computer science literature: **differential privacy**. Differential privacy is primarily studied in the context of the collection, analysis, and release of aggregate statistics. These range from simple statistical estimations, such as averages, to machine

learning. First presented in 2006 [3], differential privacy is the subject of ongoing research to develop privacy technologies that provide robust protection against a wide range of potential attacks, including types of attacks currently unforeseen. Importantly, differential privacy is not a single tool but a *definition* or *standard* for quantifying and managing privacy risks for which many technological tools have been devised. Analyses performed with differential privacy differ from standard statistical analyses, such as the calculation of averages, medians, and linear regression equations, in that random noise is added in the computation. Tools for differentially private analysis are now in early stages of implementation and use across a variety of academic, industry, and government settings.

In the following sections, we provide a simplified and informal, but mathematically accurate, description of differential privacy. Using intuitive illustrations and limited mathematical formalism, we discuss the definition of differential privacy, how it addresses privacy risks, how differentially private analyses are constructed, and how such analyses can be used in practice. We conclude with some advanced topics and pointers for further reading.

Part I
Basics of differential privacy

Consider an analysis on data containing personal information about individuals. The analysis may be as simple as determining the average age of the individuals in the data, or it may be more complex and utilize sophisticated modeling and inference techniques. In any case, the analysis involves performing a computation on input data and outputting the result. This broad notion of an analysis also includes, for example, the application of an Statistical Disclosure Limitation (SDL) technique to aggregate or de-identify a set of data, with the goal of producing a sanitized version of the data that is safe to release. In other words, we use the terms analysis and computation interchangeably to refer to any transformation, usually performed by a computer program, of input data into some output. This notion of an analysis is illustrated in Figure 1.

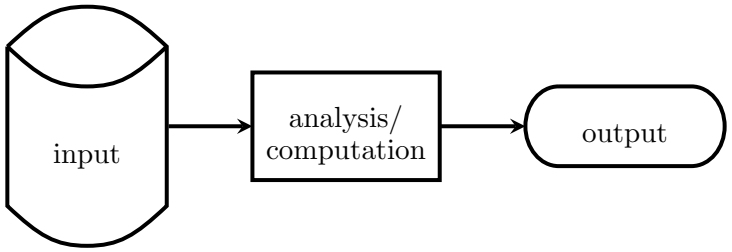

Figure 1: An analysis (or computation) transforms input data into some output.

Using this terminology, the question of whether privacy is preserved or not is not a question of whether a particular output preserves privacy but rather the question whether a particular *computation* preserves privacy: The computation that is applied to the input data determines the informational relationship between the input information and the output and hence it is the computation that we need to inspect to determine whether privacy is preserved.[4]

2 What is the differential privacy guarantee?

Intuitively, an analysis protects the privacy of individuals in the data if its output does not reveal any information about any specific individual. Differential privacy formalizes this intuition as a mathematical definition. This definition can, in turn, be used to design a privacy-preserving analysis that provides this mathematical guarantee of privacy protection. In this framework, privacy is not just a property of the output, but rather a property of the computation that generated the output.[5]

To see how differential privacy formalizes this privacy requirement, consider the following scenario.

[4] For a more in depth discussion, refer to Section 10.

[5] See Section 10 for further discussion of the importance of defining privacy as a property of the computation.

Researchers selected a sample of individuals to participate in a survey exploring the relationship between socioeconomic status and medical outcomes across a number of U.S. cities. Individual respondents were asked to complete a questionnaire covering topics such as where they live, their finances, and their medical history. One of the participants, John, is aware that individuals have been re-identified in previous releases of de-identified data and is concerned that personal information he provides about himself, such as his HIV status or annual income, could one day be revealed in de-identified data released from this study. If leaked, the personal information John provides in response to the questionnaire used in this tudy could lead to an increase in his life insurance premium or an adverse decision on a mortgage application he submits in the future.

Differential privacy can be used to address John's concerns in this scenario. If an analysis on the data from this study is designed to be differentially private, then John is guaranteed that even though his information is used in the analysis, the outcome of the analysis will not disclose anything that is *specific to him*.

To understand what this means, consider a thought experiment, which we illustrate in Figure 2 and refer to as *John's opt-out scenario*. John's opt-out scenario is one in which an analysis is performed using data about the individuals in a sample as usual, with one exception: information about John is omitted. Because John's information is omitted from the analysis, his privacy is protected in the sense that the outcome of the analysis *does not depend on his specific information*. To observe that it is indeed true that the outcome of the analysis in this scenario does not depend on John's information, note that the outcome of the analysis would not change at all if John's personal details were completely different.

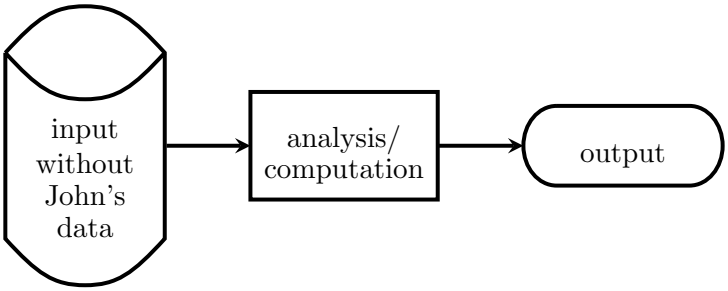

Figure 2: John's opt-out scenario.

John's opt-out scenario is distinguished from the *real-world scenario*, which involves an analysis based on John's information along with the personal information of others. The real-world scenario therefore involves some potential risk to John's privacy. Because John's information is used as input to the analysis, personal information about him could be revealed in the outcome of the analysis, though the amount of information revealed about John from such an analysis can often be quite small.

2.1 What does differential privacy protect and what does it not protect?

Differential privacy aims to protect John's privacy in the real-world scenario in a way that mimics the privacy protection he is afforded in his opt-out scenario.[6] Accordingly, what can be learned about John from a differentially private computation is (essentially) limited to what could be learned about him from everyone else's data *without his own data being included in the computation.* Crucially, this very same guarantee is made not only with respect to John, but also with respect to every other individual contributing his or her information to the analysis!

A more precise description of the differential privacy guarantee requires the use of formal mathematical language, as well as technical concepts and reasoning that are beyond the scope of this document. Rather than providing a full, precise definition, this document offers a few illustrative examples to discuss various aspects of differential privacy in a way we hope is intuitive and accessible. The examples below illustrate what is protected in real-world scenarios with and without the use of differential privacy. They also explore John's opt-out scenario in more detail. We will see that, even in John's opt-out scenario, an analysis may reveal information about John that could embarrass him, harm his social status, or adversely affect his employability or insurability in the future.

Examples illustrating what differential privacy protects

The scenarios described in this section illustrate the types of information disclosures that are controlled when using differential privacy.

> Alice and Bob are professors at State University. They both have access to a database that contains personal information about students at the university, including information related to the financial aid each student receives. Because it contains personal information, access to the database is restricted. To gain access, Alice and Bob were required to demonstrate that they planned to follow the university's protocols for handling personal data, by undergoing confidentiality training and signing data use agreements proscribing their use and disclosure of personal information obtained from the database.
>
> In March, Alice publishes an article based on the information in this database and writes that "the current freshman class at State University is made up of $3,005$ students, 202 of whom are from families earning over $\$1,000,000$ per year." Alice reasons that, because the figure in her article is an average taken over $3,005$ people, no individual's personal information will be exposed. The following month, Bob publishes a separate article containing these figures: "201 families in State University's freshman class of $3,004$ have household incomes exceeding $\$1,000,000$ per year." Neither Alice nor Bob is aware that they have both published similar information.

[6] It is important to note that the use of differentially private analyzes is *not* equivalent to the traditional use of opting out. On the privacy side, differential privacy does not require an explicit opt-out. In comparison, traditional use of opt-out require an explicit choice that may call to attention, further inspection, and hence harm to privacy. On the utility side, there is no general expectation using differential privacy would yield the same outcomes as adopting the policy of opt-out.

A clever student Eve reads both of these articles and notices the discrepancy. From the published information, Eve concludes that between March and April one freshman withdrew from State University and that the student's parents earn over $\$1,000,000$ per year. Eve asks around and is able to determine that a student named John dropped out around the end of March. Eve then informs her classmates that John's parents earn over $\$1,000,000$ per year.

John hears about this and is upset that his former classmates learned that his parents earn over $\$1,000,000$ per year. He complains to the university and Alice and Bob are asked to explain. In their defense, both Alice and Bob argue that they published only information that had been aggregated over a large population and does not identify any individuals.

This story illustrates how, in combination, the results of multiple analyses using information about the same people may enable one to draw conclusions about individuals in the data. Alice and Bob each published information that, in isolation, seems innocuous. However, when combined, the information compromised John's privacy. This type of privacy breach is difficult for Alice or Bob to prevent individually, as neither knows what information has already been revealed or will be revealed by others in future. This problem is referred to as the problem of *composition*.

Consider next what would happen if Alice and Bob had added random noise to their counts before publishing them.

Suppose, in the example above, Alice and Bob decided to add random noise to the figures they published in their articles. For the number of freshmen who come from families with a household income exceeding $\$1,000,000$, Alice publishes a count of 204 for the month of March, and Bob publishes a count of 199 for the month of April. The publication of these noisy figures would have prevented Eve from concluding that one student withdrew from the university in March and that this student came from a family with a household income exceeding $\$1,000,000$, thereby reducing the risk that John's personal information could be uncovered based on these publications.

This example hints at how differential privacy is achieved and how it addresses the problem of composition. Through the careful addition of random noise, the definition of differential privacy can be satisfied, even when the results of multiple analyses are combined. If multiple analyses are performed on data from the same set of individuals, then, as long as each of the analyses satisfies differential privacy, it is guaranteed that all of the information released, when taken together, will still be differentially private.[7]

The next example illustrates how if multiple parties both publish differentially private statistics about the same individuals, then the combination of these statistics would also be differentially private.

[7]Note that this does not mean that privacy does not degrade after multiple differentially private computations. See Section 3.2 below for a more detailed discussion of composition.

> Suppose Alice and Bob independently release statistics about the average household income of the freshman class at State University. Alice distorts the average income she intends to publish by applying a technique that satisfies differential privacy. Likewise, Bob distorts the average income he plans to publish, also using a technique that satisfies differential privacy. In doing so, without having to decide on which particular techniques to use, Alice and Bob can be sure that even in combination the information they plan to publish still satisfies differential privacy (albeit with somewhat weaker parameters than would be the case in a single release).

This example illustrates one of the greatest strengths of differential privacy: the ability to measure and bound the cumulative privacy risk from multiple analyses on information about the same individuals.

It is important to note, however, that every analysis results in some leakage of information about the individuals whose information is being analyzed and that this leakage accumulates with each analysis. This is true for every release of data, including releases of aggregate statistics, as we describe in further detail in Sections 3.2 and 8.2 below. For this reason, there is a limit to how many analyses can be performed on a specific dataset while providing an acceptable guarantee of privacy. This is why it is critical to measure privacy loss and to understand quantitatively how risk can accumulate.

Examples illustrating what differential privacy does not protect

Next, we provide examples that illustrate the types of information disclosures differential privacy does not aim to address.

> Suppose Alice is a friend of John's and possesses some knowledge about him, such as that he regularly consumes several glasses of red wine with dinner. Alice later learns of a medical research study that found a positive correlation between drinking red wine and the occurrence of a certain type of cancer. She might therefore conclude, based on the results of this study and her prior knowledge of John's drinking habits, that he has a heightened risk of developing cancer.

It may seem at first that the publication of the results from the medical research study enabled a privacy breach by Alice. After all, learning about the study's findings helped her infer new information about John that he himself may be unaware of, i.e., his elevated cancer risk. However, notice how Alice would be able to infer this information about John even if John had not participated in the medical study—i.e., it is a risk that exists in both John's opt-out scenario and the real-world scenario. In other words, this risk applies to everyone, regardless of whether they contribute personal data to the study or not.

Consider a second example:

> Alice knows that her friend John is a public school teacher with five years of experience, and he is about to start a job in a new school district. She later comes across a local news article about a teachers union dispute, which includes salary

figures for the public school teachers in John's new school district. Alice is able to determine John's new salary, based on the district's average salary for a teacher with five years of experience.

Note that, as in the previous example, Alice can determine information about John (i.e., his new salary) from the published information, even though the published information was not based on John's information. In both examples, John could be adversely affected by the discovery of the results of an analysis, even within his opt-out scenario. In both John's opt-out scenario and in a differentially-private real-world scenario, it is therefore not guaranteed that *no* information about John can be revealed. The use of differential privacy only guarantees that *no information specific to John* is revealed.

These examples suggest, more generally, that any useful analysis carries a risk of revealing some information about individuals. We argue, however, that such risks are largely unavoidable. In a world in which data about individuals are collected, analyzed, and published, John cannot expect better privacy protection than is offered by his opt-out scenario because he has no ability to prevent others from participating in a research study or a release of public records. Moreover, the types of information disclosures enabled in John's opt-out scenario often result in individual and societal benefits. For example, the discovery of a causal relationship between red wine consumption and elevated cancer risk can inform John about possible changes he could make in his habits that would likely have positive effects on his health. In addition, the publication of public school teacher salaries may be seen as playing a critical role in transparency and public policy, as it can help communities make informed decisions regarding appropriate salaries for their public employees.

2.2 How is differential privacy achieved?

One of the examples above, in which Alice and Bob add random noise to the statistics they publish in order to make it more difficult for someone to learn about an individual in the data by comparing the two sets of statistics, alludes to how differential privacy can be achieved. In order to mask the differences between a real-world computation and an individual's opt-out scenario, and thereby achieve differential privacy, an analysis must introduce some amount of *randomness*. That is, analyses performed with differential privacy differ from standard statistical analyses, such as the calculation of averages, medians, and linear regression equations, in that random noise is added in the computation. This means that the outcome of a differentially private analysis is not exact but an *approximation*, and a differentially private analysis may, if performed twice, return different results. We provide a more detailed discussion of the construction of differentially private analyses in Section 7 below.

3 The privacy loss parameter

An essential component of a differentially-private mechanism is the privacy loss parameter. For an introduction to this parameter, let us first revisit the opt-out scenario for a certain computation, such as estimating the number of HIV-positive people in a surveyed population. Ideally, this estimate should remain exactly the same whether or not a single individual, such as John, is included in the survey. However, ensuring this property *exactly* would require the total exclusion of John's information from the analysis. It would also require the exclusion of Gertrude's and

Peter's information, in order to provide privacy protection for them as well. We could continue with this argument and remove the personal information of every single surveyed individual in order to satisfy their individual opt-out scenarios. However, in doing so, we would have to conclude that the analysis cannot rely on any person's information, and hence it would be useless.

To avoid this dilemma, differential privacy requires only that the output of the analysis remain *approximately* the same, whether John participates in the survey or not. That is, differential privacy permits a slight deviation between the output of the real-world analysis and that of each individual's opt-out scenario.

A parameter quantifies and limits the extent of the deviation between the opt-out and real-world scenarios. As shown in Figure 3 below, this parameter is usually denoted by the Greek letter ϵ (epsilon) and referred to as the privacy parameter, or, more accurately, the privacy loss parameter.[8] The parameter ϵ measures the effect of each individual's information on the output of the analysis. It can also be viewed as a measure of the additional privacy risk an individual could incur beyond the risk incurred in the opt-out scenario. Note that in Figure 3 we have replaced John with a prototypical individual X to emphasize that the differential privacy guarantee is made simultaneously to *all* individuals in the sample, not just John.

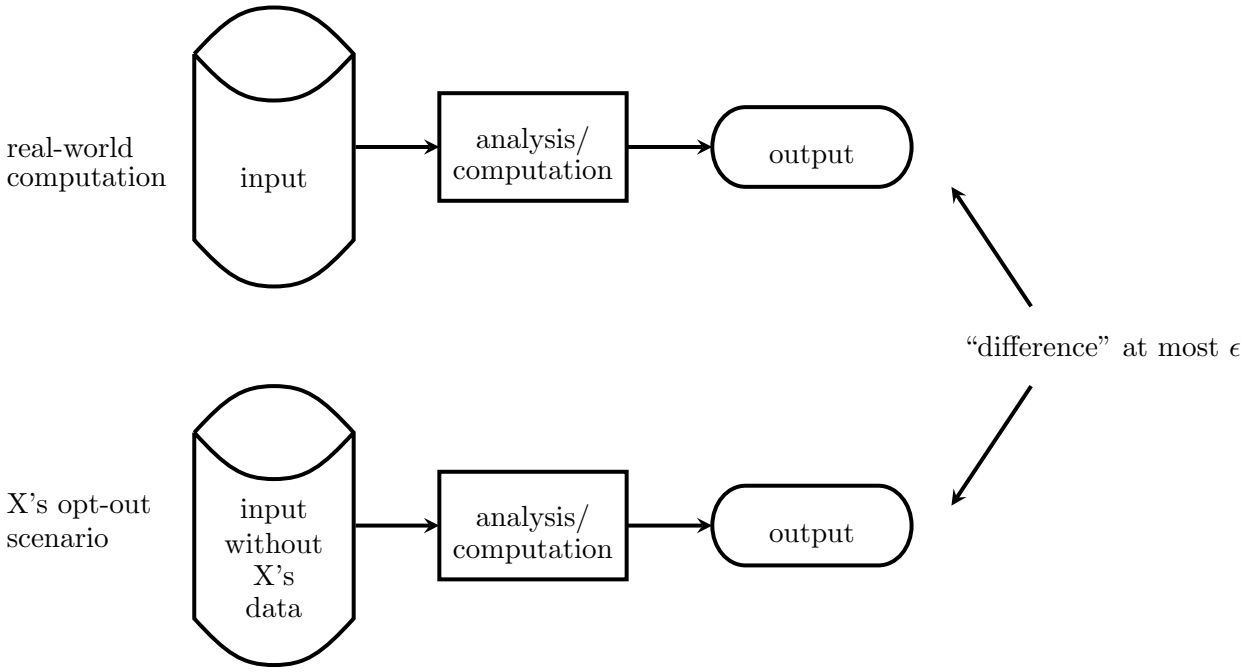

Figure 3: Differential privacy. The maximum deviation between the opt-out scenario and real-world computation should hold simultaneously for each individual X whose information is included in the input.

[8] In some implementations of differential privacy, a second parameter denoted by the Greek letter δ (delta) is also used. The parameter δ controls the probability that a privacy breach event would happen, and hence should be kept very small (e.g., one in a billion). To simplify the presentation here, we will assume that δ is set to zero.

Choosing a value for ϵ can be thought of as tuning the level of privacy protection required. This choice also affects the utility or accuracy that can be obtained from the analysis. A smaller value of ϵ results in a smaller deviation between the real-world analysis and the opt-out scenario, and is therefore associated with stronger privacy protection but less accuracy. For example, when ϵ is set to zero, the real-world differentially private analysis mimics the opt-out scenario of all individuals perfectly. However, as we argued at the beginning of this section, a simultaneous mimicking of the opt-out scenarios of all individuals in the surveyed population would require ignoring all information from the input, and hence the analysis would not provide any meaningful output. Yet when ϵ is set to a small number such as 0.1, the deviation between the real-world computation and each individual's opt-out scenario will be small, providing strong privacy protection while also enabling an analyst to derive useful statistics based on the data.

As a rule of thumb, ϵ should be thought of as a small number, between approximately 1/1000 and 1. In each implementation of differential privacy, a value of ϵ that allows a reasonable compromise between privacy and accuracy should be carefully chosen. A detailed discussion on setting the parameter ϵ, including illustrations of the nature of the tradeoff between privacy and utility associated with different values of ϵ, is provided in the sections that follow.

3.1 A technical discussion of the privacy loss parameter

We now discuss the effect of the privacy loss parameter ϵ in greater technical detail. A reader encountering this concept for the first time may choose instead to skip ahead to Section 3.2.

Any analysis that is differentially private is probabilistic in nature. The reader may be familiar with analyses performed using standard statistical software in which the outcome is deterministic, meaning that executing the same analysis on the same data produces the same results every time. In contrast, executing a differentially private analysis several times on the same data can result in different answers. This is because such analyses introduce some uncertainty into the computation in the form of random noise.[9]

The following example illustrates what we mean by the effect of the introduction of random noise into a differentially private analysis.

> Consider a differentially private analysis that approximates the fraction of HIV-positive individuals in a surveyed population. The outcome of such an analysis is a number between 0 and 1. For example, if 1.3% of the population is HIV-positive, then the output of the differentially private analysis might be, say, 0.012 or 1.2%. The reason that the differentially private analysis does not simply output the exact fraction 0.013 is that it protects privacy via the introduction of random noise.

To describe differentially privacy's use of random noise, we will rely on the notion of an *event* defined over the outcome of an analysis, a concept from probability theory.[10] An *event* in this case is simply a subset of the potential answers for the analysis. For example, we can define the

[9] For an explanation of how differentially private analyses are constructed, see Section 7 below.

[10] An *event* is a subset of the possible outcomes of a probabilistic experiment, i.e., a subset of the sample space. The probability of the event is the sum of probabilities assigned to each of the outcomes in the subset. For example, when a die is tossed, the possible outcomes, making up the sample space, are $\{1, 2, 3, 4, 5, 6\}$. If the die is fair then each of the outcomes has a probability of $\frac{1}{6}$. The event that the outcome is odd corresponds to the outcome being in the set $\{1, 3, 5\}$, and the probability of this event is $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$.

following event:

$$\mathcal{E}: \quad \text{the outcome of the analysis is between 0.1 and 0.2.}$$

Consider an analysis on some input data, for which event $\mathcal{E}$ would occur with some probability p. For an analysis on the input data excluding John's data, event $\mathcal{E}$ may occur with probability p'. The guarantee of differential privacy is that these two probabilities, p and p', are almost the same, i.e., the probability of the event $\mathcal{E}$ is similar whether John's data is included or excluded. More precisely we have that $p \leq (1+\epsilon) \cdot p'$. For instance, if the value of the privacy loss parameter ϵ is 0.01 then $1+\epsilon = 1.01$ and we get

$$p \leq 1.01 \cdot p'$$

and, similarly,

$$p' \leq 1.01 \cdot p.$$

In other words, if it happens to be that $p' = 0.1$, then we find that p is between $0.1/1.01 \approx 0.099$ and $0.1 \cdot 1.01 = 0.101$. (Note: this analysis is simplified and is accurate only for small values of ϵ.)

Differential privacy guarantees this bound on the ratios p/p' and p'/p not only for event $\mathcal{E}$ but for every event defined over the outcome of the analysis. Moreover, the privacy guarantee is made not only for John's opt-out scenario, but simultaneously for the opt-out scenario of every individual whose information is used in the analysis.

This next example illustrates these concepts in the context of a real-world scenario.

John is concerned that a potential health insurance provider will deny him coverage in the future, if it learns certain information about his health, such as his HIV-positive status, from a medical research database that health insurance providers can access via a differentially private mechanism. If the insurer bases its coverage decision with respect to John in part on information it learns via this mechanism, then its decision corresponds to an event defined over the outcome of a differentially private analysis.

For example, the insurer may believe (correctly or incorrectly) that John's HIV status is correlated with the outcome of an analysis estimating the fraction of residents in John's town who visited the hospital in the past month. The insurer may also believe (correctly or incorrectly) that John is most likely to be HIV-positive if the outcome of this analysis is a number between 0.1 and 0.2. In this case, the insurer may decide (justifiably or unjustifiably) to deny John's coverage when the following event occurs:

$$\mathcal{E}: \text{the outcome of the statistical analysis is between 0.1 and 0.2.}$$

To understand the effect of the privacy loss parameter in this scenario, it is not necessary for us to know how the insurer reached its decision. In fact, the insurer's decision may depend on multiple factors, including information it already knows about John. It is sufficient to consider that the insurer's decision corresponds to some *event* over the output of the analysis. If that is the case, it is guaranteed that the probability of John being denied coverage, based on the inclusion of information about him in the analysis, will not increase by a factor of more than

$1 + \epsilon$ compared to the scenario in which his information is not included in the analysis.

For instance, if John believes his probability of being denied insurance coverage is at most 5% if his information is not included in the medical research database accessed by the insurer via a differentially private mechanism with privacy loss parameter $\epsilon = 0.01$, then adding his information to the database can increase this probability to, at most,

$$5\% \cdot (1 + \epsilon) = 5\% \cdot 1.01 = 5.05\%.$$

Hence, with a privacy loss parameter taking a small value ($\epsilon = 0.01$, in this example) the probability that John is denied insurance coverage is almost the same, whether or not information about him appears in this medical research database.

3.2 The composition of differentially private analyses

Privacy risk accumulates with multiple analyses on an individual's data, and this is true whether or not any privacy-preserving technique is applied.[11] With differentially private analyses, the parameter ϵ quantifies how privacy risk accumulates through multiple analyses. For an illustration of the role this parameter plays in the composition of differentially private analyses, consider the following example.

Suppose information about John is contained in a medical research database that is used by a potential health insurance provider in two separate differentially private analyses. John is concerned that the results of these two analyses, when compared, could reveal private information about him such as his HIV status. For example, the potential health insurance provider could compare statistics on the number of HIV-positive residents in John's town, before and after he became a resident of the town, to determine his HIV-positive status and decide to deny him insurance coverage.

Fortunately for John, differential privacy limits the cumulative privacy loss from multiple analyses on his information. If the insurer's first differentially private analysis is performed with a privacy loss parameter of $\epsilon_1 = 0.01$, while the second utilizes a parameter of $\epsilon_2 = 0.03$, the two analyses can be viewed as a single analysis with a privacy loss parameter that is potentially larger than ϵ_1 or ϵ_2 but, at most,

$$\epsilon = \epsilon_1 + \epsilon_2 = 0.01 + 0.03 = 0.04.$$

Hence, if the probability of the potential insurer denying John insurance coverage is 5% when it is not based on an analysis including his information, it can increase

[11] We emphasize that this observation is true for *any* use of information, and, hence, for any approach to preserving privacy. It is not unique to differentially private analyses. However, the fact that the cumulative privacy risk from multiple analyses can be bounded is a distinguishing property of differential privacy.

to at most

$$5\% \cdot (1 + \epsilon) = 5\% \cdot 1.04 = 5.2\%,$$

when his information is included in both analyses. In this way, the use of differential privacy ensures that the increase in privacy risk from multiple analyses is very small.

For simplicity, this example uses a basic additive rule to compute the total degradation in the privacy loss parameter. A more advanced analysis of how privacy loss accumulates would show that the total degradation is actually smaller than suggested by this example. Research on this topic has led to the development of composition theorems for differential privacy that are beyond the scope of this document. What is significant to note for this introduction to the concept is that differential privacy provides a framework for measuring and bounding the cumulative privacy loss from multiple analyses of information about the same individuals. Although differential privacy is not the only available technique for quantifying privacy risk, one of its distinguishing features is that it is currently the only framework with quantifiable guarantees on how risk accumulates from a composition of multiple analyses.

4 How does differential privacy address privacy risks?

As explained above in Section 2.1, any useful analysis carries the risk that it will reveal information about individuals. While differential privacy cannot eliminate this risk, it can guarantee that the risk will be limited by quantitative bounds (expressed as a function of the privacy parameter ϵ). To understand the type of quantitative bound that can be guaranteed by differential privacy, consider the following example.

Gertrude, a 65-year-old woman, is considering whether to participate in a medical research study. While she can envision many potential personal and societal benefits that could result in part from her participation, she is concerned that the personal information she discloses in the course of the study could lead to an increase in her life insurance premium in the future.

For example, Gertrude is concerned that the tests she would undergo as part of the research study would reveal that she is predisposed to suffer a stroke and is significantly more likely to die in the coming year than the average person of her age and gender. If such information related to Gertrude's increased risk of morbidity and mortality is discovered by her life insurance company, it will likely increase her premium substantially.

Before she opts to participate in the study, Gertrude wishes to be assured that privacy measures are in place to ensure that her participation will have, at most, a limited effect on her life insurance premium.

4.1 A baseline: Gertrude's opt-out scenario

It is important to note that Gertrude's life insurance company may raise her premium based on something it learns from the medical research study, even if Gertrude does not herself participate in the study. The following example is provided to illustrate such a scenario.[12]

> Gertrude holds a $\$100,000$ life insurance policy. Her life insurance company has set her annual premium at $\$1,000$, i.e., 1% of $\$100,000$, based on actuarial tables that show that someone of Gertrude's age and gender has a 1% chance of dying in the next year.
>
> Suppose Gertrude opts out of participating in the medical research study. Regardless, the study reveals that coffee drinkers are more likely to suffer a stroke than non-coffee drinkers. Gertrude's life insurance company may update its assessment and conclude that, as a 65-year-old woman who drinks coffee, Gertrude has a 2% chance of dying in the next year. The company decides to increase Gertrude's annual premium from $\$1,000$ to $\$2,000$ based on the findings of the study.

In this example, the results of the study led to an increase in Gertrude's life insurance premium, even though she did not participate in the study. A potential increase of this nature is unavoidable to Gertrude because she cannot prevent other people from participating in the study. Using the terminology of Section 2 above, this type of effect is taken into account by Gertrude's insurance premium in her *opt-out scenario*.

4.2 Reasoning about Gertrude's risk

Next, we consider the increase in risk that is due to Gertrude's participation in the study.

> Suppose Gertrude decides to participate in the medical research study. Based on the results of medical tests performed on Gertrude over the course of the study, the researchers conclude that Gertrude has a 50% chance of dying from a stroke in the next year. If the data from the study were to be made available to Gertrude's insurance company, it might decide to increase her insurance premium from $\$2,000$ to more than $\$50,000$ in light of this discovery.
>
> Fortunately for Gertrude, this does not happen. Rather than releasing the full dataset from the study, the researchers release only a differentially private summary of the data they collected. If the researchers use a value of $\epsilon = 0.01$, then the insurance company's estimate of the probability that Gertrude will die in the next year can increase from 2% to at most
>
> $$2\% \cdot (1 + 0.01) = 2.02\%.$$
>
> Thus Gertrude's insurance premium can increase from $\$2,000$ to, at most, $\$2,020$. Gertrude's first-year cost of participating in the research study, in terms of a

[12] Figures in this example are based on data from Social Security Administration, Actuarial Life Table: Period Life Table, 2011, `http://www.ssa.gov/oact/STATS/table4c6.html`.

potential increase in her insurance premium, is at most \$20.

Note that this analysis above *does not* imply that the insurance company's estimate of the probability that Gertrude will die in the next year must increase as a result of her participation in the study, nor that if the estimate increases it must increase to 2.02%. What the analysis shows is that if the estimate increases it would not exceed 2.02%.

Consequently, this analysis *does* not imply that Gertrude would incur an increase in her insurance premium, or that if she would incur such an increase it would be of \$20. What is guaranteed is that, if Gertrude would incur an increase, this increase would not exceed \$20.

Gertrude may decide that the potential cost from participating in the research study, \$20, is too high and she cannot afford to participate with this value of ϵ and this level of risk. Alternatively, she may decide that it is worthwhile. Perhaps she is paid more than \$20 to participate in the study or the information she learns from the study is worth more than \$20 to her. The significance is that differential privacy allows Gertrude to make a more informed decision based on the worst-case cost of her participation in the study.

4.3 A general framework for reasoning about privacy risk

Differential privacy provides a general framework for reasoning about the increased risk that is incurred when an individual's information is included in a data analysis. Calculations like those used in the analysis of Gertrude's privacy risk can be performed by referring to Table 1. For example, the value of epsilon used in the medical research study Gertrude considered participating in was 0.01, and the baseline privacy risk in her opt-out scenario was 2%. As shown in Table 1, these values correspond to a worst-case privacy risk of 2.02% in her real-world scenario. Notice also how the calculation of risk would change with different values. For example, if the privacy risk in Gertrude's opt-out scenario were 5% rather than 2% and the value of epsilon remained the same, then the worst-case privacy risk in her real-world scenario would be 5.05%.

Note that the above calculation requires certain information that may be difficult to determine. In particular, the 2% baseline in Gertrude's opt-out scenario (i.e., Gertrude's insurer's belief about her chance of dying in the next year) is dependent on the results from the medical research study, which Gertrude does not know at the time she makes her decision whether to participate. Fortunately, differential privacy provides guarantees for every event and every baseline value.[13]

If, for example, Gertrude were to decline to participate in the study but the study results would cause Gertrude's insurer to believe that her chance of dying in the next year is 3%, then this would be the baseline for calculating that, with Gertrude's participation, the insurer's estimate for Gertrude's mortality could increase to at most $3\% \cdot (1 + 0.01) = 3.03\%$.

More generally, we can use Table 1 above to reason about how the participation of an individual in a differentially private analysis can affect the belief an insurer or any other entity may have about

[13] For the definition of event see Footnote 10 in Section 3.1.

posterior belief given $A(x')$ in %	value of ϵ					
	0.01	0.05	0.1	0.2	0.5	1
0	0	0	0	0	0	0
1	1.01	1.05	1.1	1.22	1.64	2.67
2	2.02	2.1	2.21	2.43	3.26	5.26
5	5.05	5.24	5.5	6.04	7.98	12.52
10	10.09	10.46	10.94	11.95	15.48	23.2
25	25.19	25.95	26.92	28.93	35.47	47.54
50	50.25	51.25	52.5	54.98	62.25	73.11
75	75.19	75.93	76.83	78.56	83.18	89.08
90	90.09	90.44	90.86	91.66	93.69	96.07
95	95.05	95.23	95.45	95.87	96.91	98.1
98	98.02	98.1	98.19	98.36	98.78	99.25
99	99.01	99.05	99.09	99.18	99.39	99.63
100	100	100	100	100	100	100
	maximum posterior belief given $A(x)$ in %					

Table 1: Maximal change between posterior beliefs in Gertrude's opt-out and real-world scenarios. The notation $A(x')$ refers to the application of the analysis A on the dataset x' which does not include Gertrude's information. As this table shows, the use of differential privacy provides a quantitative bound on how much one can learn about an individual from a computation.

her, as follows.

Recall that our analysis of Gertrude's privacy risk refers to the baseline belief that an insurer may have about an event concerning Gertrude. This baseline belief refers to a hypothetical scenario in which the differentially private analysis A is performed but without the individual's information taken into consideration. Denoting the data collected for the analysis by x and the same dataset without Gertrude's information by x', we refer to this hypothetical baseline belief of the insurer as the *posterior belief given* $A(x')$. The *real* belief of the insurer is formulated given the outcome of the analysis applied to the entire dataset (i.e., including Gertrude's data) $A(x)$. We call this belief the insurer's *posterior belief given* $A(x)$.

Given this terminology, we can use Table 1 to reason about the maximal difference between the belief the insurer would have had should the analysis have been performed on x' (i.e., in Gertrude's opt-out scenario) and the belief the insurer would have should the analysis be performed on x (i.e., in Gertrude's opt-in scenario). The table provides a range of posterior beliefs given $A(x')$ between zero and a hundred percent, and can be used for any potential cost Gertrude may be concerned about arising from her participation in the study. For instance, her health insurance premium (in addition to her life insurance premium) may be affected by the outcome of the study. Reasoning about each of these potential effects requires multiple, but similar, calculations.

Part II
Differential privacy and legal requirements

Social scientists and others who collect, process, analyze, store, or share data about individuals must take steps to protect the privacy of the subjects of the data in accordance with various laws, institutional policies, contracts, ethical codes, and best practices. In some settings, differential privacy can be used by researchers to analyze and share data, while both complying with such legal obligations and providing strong mathematical guarantees of privacy protection for research subjects.

To understand how differential privacy can be used to satisfy legal requirements, we must first take a closer look at legal approaches to privacy. The following scenario illustrates the types of questions that arise when using differential privacy tools to satisfy legal requirements.

> Alice and Bob, professors at State University, are co-authors of an article presenting results from a study utilizing data they obtained from a university database containing personal information about students. A journal has agreed to publish Alice and Bob's article, but it requires that they also make the data underlying the conclusions in their article available through its repository.
>
> To protect the privacy of the students in their dataset, Alice and Bob use a tool that satisfies differential privacy to produce a synthetic version of the dataset and intend to upload this synthetic dataset to the journal's repository. However, they originally obtained the data from the university database under the terms of a data use agreement prohibiting publication of the data unless the data have been de-identified in accordance with the Family Educational Rights and Privacy Act (FERPA).
>
> Alice and Bob are uncertain whether publishing the synthetic data derived from their research dataset using differential privacy would constitute a disclosure of de-identified information as defined by FERPA and would, therefore, be permissible under FERPA and their agreement with the university.

In the section below, we present some of the concepts and legal requirements for privacy protection researchers like Alice and Bob need to consider while they disseminate information from the data they collected for their study, and how, in many cases, use of differential privacy can be shown to be sufficient to satisfy these requirements.

5 Introduction to concepts used in information privacy law

Relevant legal requirements vary widely depending on the actors and institutions, the types of information, and the jurisdictions of the actors and research subjects involved. For example, the Federal Policy for the Protection of Human Subjects[14] requires researchers who receive US federal

[14] 45 C.F.R. Part 46.

funding to submit to oversight by an institutional review board (IRB) and implement informed consent and disclosure limitation procedures as directed by the IRB. Sector-specific privacy laws in the United States such as the Family Educational Rights and Privacy Act[15] and the Health Insurance Portability and Accountability Act Privacy Rule[16] prohibit the disclosure of certain types of personal information by educational institutions and health care providers, respectively, but permit the disclosure of information that has been de-identified in accordance with the standards they set forth. Laws at the state level may impose a range of additional requirements such as mandatory procedures for data security, breach notification, and data destruction.[17] Other privacy and data protection laws are in place across the globe, and have additional implications for data that cross national borders. For instance, the Data Protection Directive[18] broadly protects personal data about EU citzens and establishes rules for handling personal data within the EU. In 2018, the Directive will be superseded by the General Data Protection Regulation,[19] which will extend EU data protection law to any actor holding personal data about EU citizens. Neither the Data Protection Directive nor the General Data Protection Regulation protects information characterized as anonymous data.

In this section, we explore a number of concepts that commonly appear in information privacy law. Throughout this discussion, we refer to examples from a selection of laws such as the Family Educational Rights and Privacy Act (FERPA), the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, Title 13 of the U.S. Code,[20] and the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).[21] While specific provisions of these laws, and how they have been interpreted, are provided to illustrate how these concepts are used in the regulation of privacy, the concepts discussed underlie provisions of other laws throughout the world, making this discussion applicable beyond the scope of the subset of laws discussed.

Legal requirements relevant to issues of privacy in computation rely on an understanding of a range of different concepts, such as personally identifiable information, de-identification, linkage, inference, identification risk, expert determination, consent and opt out, and purpose and access restrictions. Each of these concepts is discussed, in turn, in relation to specific provisions of the selection of information privacy laws covered in this section, and a summary is provided at the end of this section in Table 2. While none of these concepts that appear in the law refer directly to differential privacy, we show how the concept of differential privacy can be interpreted to address these concepts while accommodating differences in how these concepts are defined.

5.1 Personally identifiable information

Personally identifiable information (PII) is a central concept used in information privacy law. Legal requirements for privacy protection often define their scope of applicability in terms of PII. For instance, legal protections typically extend only to personally identifiable information, and infor-

[15] 20 U.S.C. § 1232g; 34 C.F.R. Part 99 (2013).

[16] 45 C.F.R. Part 160 and Subparts A and E of Part 164.

[17] *See, e.g.,* 201 Code Mass. Regs. §§ 17.01 et seq.

[18] European Parliament and Council Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

[19] Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.

[20] Title 13, U.S. Code.

[21] Confidential Information Protection and Statistical Efficiency Act of 2002, Title V, Pub. L. No. 107-347, 116 Stat. 2962.

mation not considered personally identifiable is not protected. Although definitions of personally identifiable information vary significantly between laws, they are generally understood to refer to the presence of pieces of information that are linkable to the identity of an individual or to an individual's personal attributes.[22]

Consider the following example, which illustrates that a collection of records, each containing information obtained from an individual, would be considered to contain PII if it is possible to link some or all of the information in a record to the individual who contributed it.

> A researcher obtains two datasets containing records obtained from individuals. In Database A each record includes, among other information, an individual's name, Social Security number, and, in some cases, home address. The researcher notes that each of these fields can be used to directly identify or otherwise be linked to an individual. For this reason, the researcher concludes that Database A contains PII.
>
> The researcher also has another database, Database B, from which all information believed to be directly identifying, such as names, Social Security numbers, and addresses, has been suppressed. For this database, the researcher considers whether the remaining information identifies individuals indirectly. The researcher observes that records in Database B contain information such as ZIP code, date of birth, and sex.

As seen in this example, information stored about individuals may contain direct and indirect identifiers. Direct identifiers include attributes such as an individual's name, address, or Social Security number, which are used as measures of identification in many everyday situations. Indirect identifiers include attributes that (often in combination) can make an individual unique in the population. Linkage is often achieved, among other ways, by taking several attributes from a record, which, in combination, serve as indirect identifiers in that they uniquely identify a person in the population. For example, the combination of ZIP code, date of birth, and sex is unique for a large percentage of the U.S. population. Using publicly available data, such as voter registration records, it is possible to link combinations of attributes from a database to specific, named individuals.[23] Therefore, linkage of a record to an individual may be possible for both Database A and Database B, despite attempts to redact directly identifying information from Database B.

Attributes other than ZIP code, date of birth, and sex may also result in identification of individuals. For example, persistent static IP addresses and MAC addresses, which are used in Internet protocols, may be considered personally identifying information. More generally, research

[22] For a general definition of *personally identifiable information*, see, e.g., Government Accountability Office, Alternatives Exist for Enhancing Protection of Personally Identifiable Information (2008). ("For purposes of this report, the terms *personal information* and *personally identifiable information* are used interchangeably to refer to any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individuals identity, such as name, Social Security number, date and place of birth, mothers maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."). For a survey of various definitions of *personally identifiable information*, see Paul M. Schwartz & Daniel J. Solove, The PII Problem: Privacy and a New Concept of Personally Identifiable Information, 86 N.Y.U. L. Rev. 1814 (2011).

[23] *See* Latanya Sweeney, *Simple Demographics Often Identify People Uniquely*, Data Privacy Lab Technical Report (2000).

has shown that combinations of attributes that are not usually thought of identifying can often lead to identification of individuals.[24]

Due in large part to concerns such as these, laws have been enacted with the aim of protecting types of information that could be used to identify individuals in a database. For instance, FERPA protects *non-directory personally identifiable information* maintained by educational agencies and institutions. FERPA defines PII by way of a non-exhaustive list of examples of information considered directly or indirectly identifying, as well as "[o]ther information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty."[25] FERPA distinguishes between non-directory PII and *directory information* which may include names, phone numbers, and other information that the institution chooses to include and "would not generally be considered harmful or an invasion of privacy if disclosed."[26] The HIPAA Privacy Rule regulates the use and disclosure of *protected health information* by covered health care entities and their business associates. Protected health information is defined as individually identifiable health information, which "identifies the individual" or "with respect to which there is a reasonable basis to believe the information can be used to identify the individual."[27] Title 13 of the U.S. Code governs the U.S. Census Bureau's activities, including its confidentiality requirements when handling information supplied by respondents participating in its censuses and surveys. The statute does not explicitly refer to PII but to "any publication whereby the data furnished by any particular establishment or individual . . . can be identified."[28] CIPSEA establishes confidentiality protection for information collected by federal agencies directly from respondents under a pledge of confidentiality for exclusively statistical purposes. Its protections apply to information in *identifiable form*, defined as "any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means."[29]

The term *personally identifiable information* does not have a precise technical meaning, and in practice it can be difficult to determine whether information is personal, identifying, or likely to be considered identifying in the future. Indeed, definitions between laws and the interpretations of these definitions vary widely. There is no clear rule for determining whether particular pieces of information are personally identifying, or for determining whether an information release should be considered to contain personally identifiable information. Experts often disagree whether a particular piece of information constitutes personally identifiable information, and it is a notion that is evolving as practices adapt in response to new privacy attacks. Over time, successful demonstrations of privacy attacks reveal new privacy vulnerabilities that are arguably not addressed by existing interpretations of personally identifiable information. Furthermore, determining what is personally identifiable information in releases that are not in a microdata or tabular format, such as statistical models or outputs of a machine learning system, is unclear.

Regardless of the definition or notion of personally identifiable information that is used, differen-

[24] Narayanan and Shmatikov generalize these findings and assert that "[a]ny information that distinguishes one person from another can be used for re-identifying anonymous data." *See* Arvind Narayanan & Vitaly Shmatikov, *Myths and Fallacies of "Personally Identifiable Information,"* 53 Communications of the ACM 24, 26 (2010).

[25] 34 C.F.R. § 99.3.

[26] 34 C.F.R. § 99.3.

[27] *See* 45 C.F.R. § 160.103.

[28] 13 U.S.C. § 9.

[29] Pub. L. 107-347 § 502(4).

tial privacy can be interpreted as ensuring that using an individuals data will not reveal personally identifiable information specific to her.[30] Here, *specific* refers to information that is unique to the individual and cannot be inferred unless the individual's information is used in the analysis. This interpretation of the differential privacy guarantee can be used to argue that the use of differential privacy is sufficient to satisfy the requirements of information privacy laws that protect personally identifiable information from disclosure.

5.2 De-identification

The term *de-identification* refers to a collection of techniques that aim to transform identifiable information into non-identifiable information, while also preserving some utility of the data. In principle, it is intended that de-identification, *if performed successfully*, can be used as a tool for removing PII, or transforming PII into non-PII.

A host of techniques and best practices for de-identification have been developed over the years and are currently used to transform PII into non-PII with varying degrees of success. The statistical disclosure limitation literature includes a wide range of techniques for suppression, noise addition, swapping of information among records, and synthetic data generation that are used for de-identification. For example, tools to *suppress* direct or indirect identifiers are applied with the goal of preventing linkage of a record with other available datasets. When using suppression techniques to de-identify information, it is generally recommended to take into consideration what data may be available to a privacy attacker presently or in the future. These techniques can provide protection against certain types of attacks. However, use of suppression techniques alone is widely considered to be insufficient to protect privacy, and, as a result, such techniques are commonly used in combination with other methods. Another technique commonly used in the publication of statistical tables is the suppression of cells representing small groups. The rationale underlying the suppression of small cell counts is that cells with low counts (such as counts between 1 and 4) indicate the existence of uniques in the population or sample, and can lead to identification, or learning of sensitive attributes of individuals.

A number of information privacy laws explicitly or implicitly authorize the release of information about individuals after it has been de-identified. For instance, by definition, data that have been de-identified according to the FERPA standard are considered not to contain any personally identifiable information and can be used for any purpose without restriction.[31] The Department of Education declines, however, to prescribe specific methods that are sufficient for redacting records or producing statistical information in accordance with these requirements, explaining that "determining whether a particular set of methods for de-identifying data and limiting disclosure risk is adequate cannot be made without examining the underlying data sets, other data that have been released, publicly available directories, and other data that are linked or linkable to the information in question."[32] Instead, the agency provides a list of examples of statistical disclosure limitation methods that may be used, such as adhering to a minimum cell size, controlled rounding, top coding, replacing individual values with categorical groupings, and data swapping, and refers practitioners to a summary of statistical disclosure limitation techniques prepared by the Federal Committee on Statistical Methodology for additional guidance.[33]

[30] Note that the term *use* in this statement refers to the inclusion of an individual's data in an analysis.
[31] *See* 73 Fed. Reg. 74,806, 74,836.
[32] *Id.*
[33] *Id.*; 73 Fed. Reg. 15,574, 15,584 (Mar. 24, 2008).

The HIPAA Privacy Rule establishes a de-identification standard that is satisfied when protected health information is transformed such that (1) the health information does not identify an individual and (2) with respect to which there is no reasonable basis to believe that the information can be used to identify an individual.[34] There are two methods for de-identifying information in accordance with the Privacy Rule: the expert determination method and the safe harbor method.[35] The safe harbor method involves the removal of eighteen pieces of information considered to be identifying, including names, telephone numbers, Social Security numbers, and medical record numbers, among others.[36] In addition, the covered entity must not have "actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information."[37] De-identification in accordance with the second method, expert determination, is satisfied by an expert applying "generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" and determining that "the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information."[38]

Title 13 and CIPSEA are typically interpreted to require the application of statistical disclosure limitation techniques, such as cell suppression, data swapping, rounding, top and bottom coding, random noise addition, and synthetic data generation.[39] Such techniques are used to transform information prior to release, with the goal of preventing the identification of respondents who supplied data for statistical purposes.

Generally, de-identification is intended as a means to transform PII into non-PII. In other words, de-identification techniques are used by organizations that aim to disclose information that does not contain personal information that can be linked to specific individuals whose information has been used. As discussed in the previous section, any algorithm that satisfies the requirements of differential privacy has the property that using an individuals data will not reveal personally identifiable information specific to her. Because the output of a differentially private computation does not reveal personally identifiable information, any differentially private algorithm should be considered sufficient for de-identification. An alternative interpretation is that differential privacy provides very strong privacy protection for individuals, substantially reducing the risk that including information about an individual in an analysis will reveal information that is specific to them, including their identity or sensitive attributes.

Moreover, differential privacy can provide some utility where traditional approaches to privacy do not. In particular, differential privacy can be used to safely compute some statistics using information considered to be PII. For example, differential privacy can be used to safely generate statistics on the relationship between an individual's first name and lifetime earnings, whereas a de-identified dataset that has been stripped of individuals' names could not support such an analysis.

[34] *See* 45 C.F.R. § 164.514(a).
[35] *See* 45 C.F.R. § 164.514(b).
[36] 45 C.F.R. § 164.514(b)(2).
[37] § 164.514(b)(2)(ii).
[38] 45 C.F.R. § 164.514(b)(1).
[39] *See, e.g.,* Amy Lauger, Billy Wisniewski, & Laura McKenna, *Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research*, U.S. Census Bureau Research Report (2014).

5.3 Linkage

Guidance on complying with legal requirements for privacy protection often refers to specific modes of privacy loss. References to these modes of privacy loss are intended to highlight certain types of privacy attacks that should be addressed by the measures implemented by privacy practitioners. Among the various modes of privacy loss frequently referenced by privacy regulations are linkage (often using auxiliary information), singling out, and inference. With the discussion in this and the following section, we explain how differential privacy provides privacy protection against a very large class of modes of privacy loss, including those commonly referenced within legal guidance.

One of the most prominent modes of privacy loss recognized by privacy regulations, implicitly or explicitly, is a successful record linkage. Linkage typically refers to the matching of information in a database to a specific individual, often by leveraging auxiliary data sources. For example, by defining personally identifiable information in terms of information "linked or linkable to a specific student,"[40] FERPA appears to emphasize the risk of a successful record linkage attack. The Department of Health & Human Services in guidance on de-identifying data in accordance with the HIPAA Privacy Rule includes an extended discussion of examples of record linkage attacks and de-identification strategies for mitigating them.[41]

Linkage is also closely related to the concept of *identifying* an individual in a data release, as identifying an individual is often accomplished via a successful linkage. Another variant of the concept of linkage referenced in EU data protection law, including the new General Data Protection Regulation, refers to *singling out*, which "corresponds to the possibility to isolate some or all records which identify an individual in the dataset,"[42] and appears to focus on data releases in microdata formats, in which each record corresponds to an individual.

Attacks on de-identified data often leverage external, or auxiliary, information, which can be used to link de-identified information to personally identifiable information. Examples of auxiliary information include publicly available information, such as information from a voter registration record or a social media profile, or personal knowledge that a friend, relative, or colleague might have about an individual in a database. For instance, as mentioned above, it has been shown that an individual's birth date, ZIP code, and sex can be retrieved from a voter registration record, and used to link an individual's identity to a record in a de-identified database containing the same three attributes. Similarly, it was demonstrated that individuals could be identified in Netflix's database of users' movie ratings using another source of individuals' movie ratings, such as the Internet Movie Database, or even limited knowledge about a friend's movie viewing history.[43] Concerns about attacks such as these are reflected in regulatory guidance on complying with legal requirements, which often advise that, when applying de-identification techniques, practitioners should consider the data sources that are available and could be leveraged in a linkage attack.

Linkage attacks have a concrete meaning when data is published as a collection of individual-level records, often referred to as microdata. However, what is considered a successful linkage when a publication is made in other formats (including, e.g., statistical models and synthetic data) is open to interpretation. Despite this ambiguity, it can be argued that differential privacy addresses

[40] 34 C.F.R. § 99.3.

[41] *See* Office for Civil Rights, Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* (2012).

[42] *See* Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques (2014).

[43] *See* Arvind Narayanan & Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, IEEE Security and Privacy (2008).

any reasonable interpretation of record linkage. For example, microdata or contingency tables that allow the identification of population uniques cannot be created using statistics produced by a differentially private tool, making it structurally impossible to isolate or link an individual's record given a statistical output generated using differential privacy. Similarly, differential privacy masks the contribution of a single individual, making it impossible to infer any information specific to an individual, including whether an individual's information was used at all. Differentially private statistics also provably hide the influence of every individual, and even groups of individuals, providing protection not only against releasing exact records but also approximate statistics that could leak individual-level information. Furthermore, differential privacy provides a robust guarantee of privacy protection that is independent of the auxiliary information available to an attacker. When differential privacy is used, an attacker utilizing auxiliary information cannot learn much more about an individual in a database than she could if that individual's information were not in the database at all. It follows from these interpretations that an individual's record cannot be singled out or linked using the statistics produced by a differentially private tool.

5.4 Inference

Some information privacy laws, or interpretations of these laws, refer to modes of privacy loss involving *inference*. Some laws refer to inference explicitly. For example, CIPSEA protects "any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means."[44] EU data protection law also refers to inference, defined as "the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes."[45] Other laws are interpreted to prohibit disclosures of information that enable one to determine an attribute about an individual with (high) certainty. For example, FERPA defines *personally identifiable information*, in part, in terms of information that would allow one to identify a student "with reasonable certainty."[46] In practice, FERPA has been interpreted to require the use of statistical disclosure limitation techniques for preserving uncertainty, such as using minimum group sizes, withholding exact counts in favor of rounded percentages, and applying top and bottom coding procedures to percentages prior to release.[47]

When discussing inference, it is important to distinguish between two types: inferences about individuals and inferences about large groups of individuals. The law generally does not draw a clear distinction between these two types of inference. However, differentiating between these two categories of inference is key to enabling socially beneficial uses of data, such as research investigating the relationship between smoking and lung cancer, while protecting individuals from disclosures of information specific to them. In explicit recognition of this distinction, differential privacy rules out inferences about individuals, thereby protecting individuals from inferences about values or attributes that are specific to them. As described above in Section 4, by protecting against such inferences, differential privacy formally bounds the increase in risk that is incurred when an individual's information is included in a data analysis. If a successful record linkage is interpreted as a reduction in the uncertainty an attacker has about an individuals data, then differential privacy

[44] Pub. L. 107-347 § 502(4) (emphasis added).

[45] *See* Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques (2014).

[46] 34 C.F.R. § 99.3.

[47] *See* National Center for Education Statistics, Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting (2011).

provides guarantees regarding protection against record linkage.

In summary, differential privacy provides privacy protection against a very large class of modes of privacy loss, including those commonly referenced within legal guidance, such as linkage, singling out, and inference.

5.5 Identification risk

Some information privacy laws refer to an acceptable level of risk of identification of a record in a data release. Similarly, other laws often acknowledge, implicitly or explicitly, that any disclosure of information carries privacy risks, and therefore the goal is to minimize rather than eliminate such risks. For example, in clarifying guidance on FERPA, the Department of Education refers to the law's de-identification requirements in terms of the goal of "minimiz[ing] the risk of disclosing personally identifiable information in redacted records or statistical information."[48] The HIPAA Privacy Rule requires covered entities to use de-identification techniques prior to releasing data in order to create a dataset with only a "very small" risk of identification.[49] Guidance on protecting confidentiality in accordance with CIPSEA requires agencies to "collect and handle confidential information to minimize risk of disclosure," among other requirements.[50] The Census Bureau's interpretation of Title 13's risk tolerance is reflected in its disclosure review policies which recognize the risks inherently associated with any publication of data and, accordingly, require a careful balancing of utility and privacy as part of each data release decision.[51]

In practice, it is not clear how to measure identification risk, as approaches vary and are largely ad hoc. Some de-identification experts aim to measure the number or percentage of records in a dataset that are likely to be identified. Another approach is to measure how one's uncertainty with respect to an individual's personal information can decrease in light of what is learned from a data publication.

Differential privacy enables a formal quantification of risk, and the privacy loss parameter epsilon can be tuned to different legal requirements for minimizing risk. Regardless of how identification risk—or privacy risk, more generally—is defined, differential privacy guarantees that the risk to

[48] 73 Fed. Reg. 74,806, 74,835 (Dec. 9, 2008).

[49] The expert determination method for de-identifying information in accordance with the HIPAA Privacy Rule requires an expert to make a determination that "the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information." 45 C.F.R. § 164.514(b)(1). The goal of the HIPAA safe harbor method has also been interpreted to be achieving a "very small risk of a privacy violation. *See* 65 Fed. Reg. 82,462, 82,543 ("The intent of the safe harbor is to provide a means to produce some de-identified information that could be used for many purposes with a very small risk of privacy violation."). In guidance, the Department of Health & Human Services recognizes that "[b]oth methods, even when properly applied, yield de-identified data that retains some risk of identification. Although the risk is very small, it is not zero, and there is a possibility that de-identified data could be linked back to the identity of the patient to which it corresponds." Office for Civil Rights, Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* (2012).

[50] *See* Office of Management and Budget, Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), 72 Fed. Reg. 33,361 (2006).

[51] *See* George Gatewood, *Census Confidentiality and Privacy: 1790 - 2002* (2001) ("[T]he Census Bureau uses [disclosure limitation] to hinder anyone from identifying an individual respondent or establishment by analyzing published census or survey data, especially, by manipulating the arithmetical relationships among the data. At the same time, the agency has the responsibility of releasing data for the purpose of statistical analysis. The desire then is to release as much statistically valid and useful data as possible without violating the confidentiality of the data as required by title 13.").

an individual is almost the same with or without her participation in the dataset. In this way, differential privacy can be interpreted to guarantee that the risk to an individual is minimal or very small.

5.6 Expert determination

Some regulations allow for an expert to make a determination regarding the appropriate disclosure limitation techniques to apply to a dataset prior to release. Notably, the HIPAA Privacy Rule provides an expert determination method by which information can be de-identified in accordance with the regulation's requirements. De-identification under the expert determination method is satisfied by an expert applying "generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" and determining that "the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information."[52] In interpreting the expert determination requirements, the Department of Health & Human Services has declined to establish required qualifications for experts or clarification of how to determine whether a de-identified information release carries "very small" risk.[53]

Statistical agencies governed by Title 13 or CIPSEA defer to the expertise of a disclosure review board on applying disclosure limitation techniques when publishing statistics based on information from respondents. Although expert determination is not required by FERPA, the Department of Education has encouraged educational agencies and institutions to consult with an expert in de-identification where feasible.[54]

In contexts in which expert determination is required (or recommended), it is possible to demonstrate that the use of differential privacy achieves what is intended by many legal requirements. As explained above, differential privacy ensures that using an individuals data will not reveal personally identifiable information specific to her. Differential privacy guarantees that the risk to an individual—including her identification risk or privacy risk, more generally—is almost the same with or without her participation in the dataset. Differential privacy is a formal privacy definition that provides general protection against a wide range of attacks, and it can be formally analyzed and tuned in accordance with the context in which it is applied.

Moreover, even where expert determination is not required by law, we believe that the use of privacy-preserving technologies should be carefully analyzed and supported by an argument that is rigorous from both a technical and a legal standpoint, and the formal mathematical definition of differential privacy is amenable to such an analysis.[55]

[52] 45 C.F.R. § 164.514(b)(1).

[53] "There is no specific professional degree or certification program for designating who is an expert at rendering health information de-identified. Relevant expertise may be gained through various routes of education and experience." *See* Office for Civil Rights, Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* (2012). "There is no explicit numerical level of identification risk that is deemed to universally meet the very small level indicated by the method. The ability of a recipient of information to identify an individual (i.e., subject of the information) is dependent on many factors, which an expert will need to take into account while assessing the risk from a data set." *See id.*

[54] In guidance, the Department of Education notes that it "recognizes that there are some practices from the existing professional literature on disclosure limitation that can assist covered entities in developing a sound approach to de-identifying data for release, particularly when consultation with professional statisticians with experience in disclosure limitation methods is not feasible." 73 Fed. Reg. 15,574, 15,584 (Mar. 24, 2008).

[55] For an example of a combined legal-technical argument that a technology satisfies a legal requirement for privacy

5.7 Consent and opt out provisions

Some information privacy laws include consent provisions, or opt out provisions, by which individuals can choose to allow, or not to allow, their information to be used by or redisclosed to a third party, respectively. For example, FERPA includes a provision requiring educational agencies and institutions to offer students an opportunity to opt out of the disclosure of their personal information in school directories.[56] Underlying consent or opt out provisions such as these are assumptions that providing individuals with an opportunity to opt in or out gives them control over the use of their personal information and effectively protects their privacy. However, these assumptions warrant a closer look. For instance, providing consent or opt out mechanisms as a means of providing individuals with greater control over their information is an incomplete solution if individuals are not also fully informed about the consequences of uses or disclosures of their information. In addition, allowing individuals the choice to opt in or out can create new privacy concerns. An individual's decision to opt out may (often unintentionally) be reflected in a data release or analysis and invite scrutiny into whether the choice to opt out was motivated by the need to hide compromising information.[57]

Differential privacy can be viewed as automatically providing all individuals in the data with the protection that opting out is intended to provide. When differential privacy is used, the consequences for an individual's privacy are almost the same whether or not an individual's information is included in an analysis. Moreover, differential privacy provides all individuals with this privacy guarantee, thereby avoiding the possibility that individuals who choose to opt out would, by doing so, inadvertently reveal a sensitive attribute about themselves or attract attention as individuals who are potentially hiding sensitive facts about themselves.

5.8 Purpose and access restrictions

Information privacy laws often include provisions restricting the use or disclosure of personally identifiable information to specific parties or for specific purposes, with wide variations in such requirements across different laws. For instance, laws governing confidentiality requirements for statistical agencies generally restrict uses of identifiable information from respondents to uses for statistical purposes. Title 13 restricts the use of confidential information from respondents, prohibiting uses "for any purpose other than the statistical purposes for which it is supplied,"[58] and restricting access to agency employees and approved researchers with Special Sworn Status. CIPSEA prohibits the use of protected information "for any use other than an exclusively statistical purpose,"[59] where *statistical purpose* "means the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups."[60] FERPA's standard for de-identifying education records applies only to releases "to any

protection, see Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R. O'Brien, Thomas Steinke, & Salil Vadhan, Bridging the Gap between Computer Science and Legal Approaches to Privacy, Working Paper (2017).

[56] 34 C.F.R. § 99.37.

[57] For a real-world example, consider recent reports that the National Security Agency's surveillance efforts specially target users of privacy services. See Kim Zetter, The NSA Is Targeting Users of Privacy Services, Leaked Code Shows, Wired, July 3, 2014.

[58] 13 U.S.C. § 9(a)(1).

[59] Pub. L. 107-347 § 512(b)(1).

[60] Pub. L. 107-347 § 502(9).

party for any purpose."[61] In accordance with FERPA, personally identifiable information can be shared with school officials with a legitimate educational interest in the information,[62] authorized representatives of the Comptroller General of the U.S., the Attorney General, the Secretary of Education, and State or local educational authorities,[63] and organizations conducting studies for, or on behalf of, schools, school districts, or postsecondary institutions.[64]

Legal requirements reflecting purpose and access restrictions such as these can be divided into two categories. Restrictions limiting use to statistical purposes, including statistical purposes involving population-level rather than individual-level analyses or statistical computations, are consistent with the use of differential privacy. Tools that satisfy differential privacy can be understood to restrict uses to only those that are for statistical purposes. However, other use and access restrictions such as restrictions limiting use to "legitimate educational purposes" are orthogonal to differential privacy.

Legal standards for privacy protection, and interpretations of these standards set forth in guidance, do not directly address the question of whether use of technologies relying on differential privacy is sufficient to satisfy their requirements. Determining whether differentially private analyses can satisfy legal requirements for protecting privacy is challenging for a number of reasons. Because privacy laws are often sector-, jurisdiction-, and context-specific, different legal requirements apply depending on the setting, leading to different requirements for various datasets held by a single institution, or different requirements for the same or similar datasets held by different institutions. In addition, many legal standards for privacy protection are, to a large extent, open to interpretation and therefore require a case-specific legal analysis by an attorney. Other challenges arise from the fact that the privacy concepts found in legal standards differ significantly from those underlying differential privacy. For instance, many laws focus on the presence of "personally identifiable information" or the ability to "identify" an individual's personal information in a release of records. Such concepts are not precisely defined, and they do not perfectly match the definition of differential privacy. Many laws also emphasize requirements for protecting privacy when disclosing individual-level data, but lack clear guidance for disclosing privacy-preserving aggregate statistics. While in some cases it may be clear whether a legal standard has been met by the use of differential privacy, in other cases—particularly along the boundaries of a standard—there may be considerable uncertainty.

Despite the conceptual gaps between differential privacy and legal requirements for privacy protection, there are strong reasons to believe that differential privacy provides privacy protection that is consistent with many legal requirements. Tools relying on differential privacy provide robust privacy protection for individuals. The strength of this approach arises from the fundamental features of differential privacy. Differential privacy is formally defined and enables a formal analysis of risk to an individual. Differential privacy provides a mathematical quantification of the excess risk to an individual from participating in an analysis, and preserves privacy taking into account the accumulation of risk over multiple analyses. The use of differential privacy prevents an adversary from determining whether a given individual's personal information was included in an analysis. In fact, it provides strong protection against a wide range of both known and unforeseeable attacks, including the types of record linkage attacks referenced, explicitly or implicitly, in the design of

[61] 73 Fed. Reg. 74,806, 74,836.
[62] *See* 34 C.F.R. §§ 99.31(a)(1), 99.7(a)(3)(iii).
[63] *See* 34 C.F.R. §§ 99.31(a)(3), 99.35.
[64] *See* 34 C.F.R. § 99.31(a)(6).

	De-identification Requirements				Purpose & Access Restrictions	
	Identification Risk Tolerance	**Scope of Attacks**	**De-identification Transformation**	**Expert Determination**	**Purpose Restrictions**	**Access Restrictions**
FERPA	Minimize risk	Record linkage, Other	Application of SDL techniques (esp. suppression, minimum cell sizes)	Recommended, but not required	Studies, audit, and legitimate educational purposes	Researchers, Authorized representatives, School officials
HIPAA Safe Harbor	Very small risk	Record linkage	Removal of 18 identifiers	Not required	Research uses with no more than minimal risk	Researchers
HIPAA Expert Determination	Very small risk	Record linkage	Application of SDL techniques (esp. k-anonymity)	Expert Analysis, Documentation	Research uses with no more than minimal risk	Researchers
Title 13	Ostensibly none (but implicitly balanced with utility)		Application of SDL techniques	Disclosure Review Board	Statistical purposes for which it is supplied	Special Sworn Status
CIPSEA	Minimize risk		Application of SDL techniques	Disclosure Review	Statistical purposes (defined in part in terms of population-level analysis)	Officers, employees, or agents of the agency

Table 2: Aspects of select requirements from information privacy law.

numerous privacy laws.

Under any one of these interpretations, it is therefore likely that in many cases a differentially private mechanism would prevent the types of disclosures of personal information that legal protections have been designed to address. In many cases, differentially private tools provides privacy protection that is more robust than techniques commonly used to satisfy legal requirements for privacy protection. However, research exploring methods for proving that differential privacy satisfies legal requirements and for tuning the privacy loss parameter ϵ based on legal requirements is needed.[65] In practice, data managers should consult with legal counsel in considering whether differential privacy tools, potentially in combination with other tools for protecting privacy and security, are appropriate within their institutional settings.

6 Implications of differential privacy for the future regulation of privacy

Differential privacy is a newly emerging concept and is currently in early stages of deployment and use. Legal scholars, policymakers, and regulators are beginning to study the definition and examine how a formal privacy model like differential privacy can be integrated with legal, policy, and regulatory approaches.[66] In this section, we review some potential implications that differential privacy—and more generally, formal privacy models—may have on privacy regulations.

Current privacy law largely reflects the traditional role of law as constraining behavior through threat of sanction. Traditional approaches to protecting privacy under the law have similarly focused on placing specific constraints on information release. However, traditional approaches are generally ad hoc, lacking a broad theoretical foundation and mathematical rigor, and having limited applicability. These factors have contributed to best practices and standards that are increasingly shown not to provide sufficient privacy protection over the long term. The line of research that yielded differential privacy has demonstrated that privacy can be reasoned about with the benefits that come from mathematical rigor, such as well-defined privacy desiderata, provable quantifiable limits on privacy risks, composition properties, a programmable framework for developing algorithms for increasingly complex task, and relationships with other research areas such as statistics, machine learning, and economics. These lessons can inform policymakers considering revisions to the legal framework in the big data era. In the discussion below, we outline how these design goals can be incorporated in the development of future privacy regulations.

Well-defined privacy goals. Many existing privacy regulations do not specify an explicit privacy goal, i.e., the desired consequences of privacy protection, known more formally as *privacy desiderata*. Rather, they describe privacy protection implicitly as a byproduct of applying a specific technique, such as de-identification via the suppression of personally identifiable information.

[65] For an extended discussion of the gaps between legal and computer science definitions of privacy and a demonstration that differential privacy can be used to satisfy an institution's obligations under the Family Educational Rights and Privacy Act, see Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R. O'Brien, Thomas Steinke, & Salil Vadhan, Bridging the Gap between Computer Science and Legal Approaches to Privacy, Working Paper (2017).

[66] *See, e.g.,* Anne Klinefelter & Andrew Chin, Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study, 90 N.C. L. Rev. 1417 (2012); Urs Gasser, Recoding Privacy Law: Reflections on the Future Relationship Among Law, Technology, and Privacy, 130 Harv. L. Rev. F. 61 (2016).

This approach reflects the traditional role of law as constraint and results in uncertainty with respect to the regulators' intended goal, particularly with respect to the use of privacy protection technologies that differ from those presented in regulations. This problem is exacerbated as the data landscape changes, computing power increases, and new privacy attacks emerge, as these factors increase uncertainty surrounding implicit privacy standards. It is unlikely that regulators intended the privacy standards they created to become weaker over time in light of technological developments, but without clearly defined privacy goals, current standards leave room for a significant degree of ambiguity and potentially create moving targets for privacy practitioners.

In contrast, differential privacy is not a specific technique but rather a definition, a privacy standard that accurately expresses a specific desideratum and can be shown to entail specific consequences to privacy risk (as discussed in detail above in Section 4). Expressing privacy in terms of these desiderata enables the use of privacy technologies as part of an adaptive *solution* to managing privacy risks. In the future, policymakers could choose to follow suit and standardize a goal, or several goals, for the regulation of privacy. Defining these goals could involve specification of basic terms used in legal standards, including precise definitions for terminology such as *personally identifiable information*, *identification*, and *linkage*, clear delineations of what constitutes an inference attack, and an explanation of the factors should be taken into account when evaluating the level of acceptable risk, among other concepts. As with differential privacy, clear privacy goals would guide the development of new technologies and ensure that future regulations are compatible with a wide range of technological solutions that can be shown to agree with them.

Quantitative measures of privacy. Quantitative measures are essential for robust privacy protection as well as informed public discussion of privacy risks. Quantitative measures of privacy could help regulators and practitioners set concrete, meaningful privacy goals and make adjustments according to the level of privacy protection required in different contexts. Incorporating a quantitative basis would also make future data releases robust to new, currently unforeseen, uses of data. Concrete quantitative measures could also help make the regulation technology-neutral by setting clear criteria for the adoption of new technologies.

Composition awareness. Regulations could be designed to require addressing the accumulated risk to privacy due to multiple uses of individual data. In addition, new legal-technological tools could be developed to limit the cumulative effect on individual privacy from multiple computations over data.

Generality of privacy protection. Many current regulations explicitly or implicitly endorse de-identification techniques like suppression of personally identifiable information as a measure that is sufficient to protect privacy. However, the failure of traditional techniques to provide adequate privacy protection has demonstrated that regulation should be designed not to endorse one specific technique but should establish a privacy goal that many techniques could be designed to satisfy. Moreover, the attack modes contemplated by the regulations seemingly address a limited family of attacks. Technological advances are leading to the emergence of new and sophisticated attacks, such as statistical inference attacks, that were unforeseen by regulators at the time that the rules were drafted. In contrast to some privacy regulations that seemingly have a narrow range of attacks in mind, the computer science literature aims to define a very broad family of attackers and prove security with respect to the entire family. Computer scientists recognize the need to protect not

only against known modes of attack, but also against unknown future attacks. Likewise, privacy regulations could be drafted to conceptualize a wider range of privacy threats, both known and unknown.

Universal coverage. The vast array of privacy regulations governing data releases make it difficult for a practitioner to assess whether application of a given privacy technique is adequate to satisfy applicable legal requirements. Privacy regulations could be designed to provide universal coverage, much like formal privacy models which are broadly applicable regardless of the type of information involved, leading to greater consistency in regulation across jurisdictions and sectors. This design goal of universal coverage is similar to the approach adopted by the General Data Protection Regulation.

Transparency in using data. Differential privacy provides a quantifiable measure of privacy loss, embodied as the privacy parameter ϵ, that can be disclosed to and understood by various stakeholders, such as data subjects and disclosure review boards. Regulators could require engineers to clearly disclose the level of privacy protection afforded by the privacy-preserving tools they implement or use. Such disclosures would enable data subjects and disclosure review boards to assess the privacy risk from a single analysis, as well as the overall risk to privacy from multiple analyses by many organizations. Unlike current privacy policies or terms of use which often fail to provide individuals with usable information about the risks they may incur by sharing their data, a disclosure of the value of ϵ an organization uses would be standardized, informative, and operable from the viewpoint of a data subject or disclosure review board.

Corporations like Google and Apple are demonstrating the feasibility of using differentially private computations for analyzing usage patterns of their users, increasing the privacy community's understanding of use cases for which differential privacy provides sufficient utility. In use cases such as these, where the goal is to learn behavior patterns about an entire user community but not about individual users, the use of differentially private analyses may evolve to become a best practice, particularly in light of known weaknesses of traditional techniques, and the reasonable anticipation that more weaknesses will be found as research progresses. This understanding could, in turn, be reflected in future regulations. Moving forward, regulators may choose to incorporate modern privacy concepts, including some of the design criteria outlined above, into revisions of existing regulations, or new regulations that specify precise privacy goals or simply establish a safe harbor for formal privacy models like differential privacy and similar approaches. Further research and discussion is needed to examine the suitability of future revisions to the regulatory framework among these possibilities.

Part III
Using differential privacy

7 How are differentially private analyses constructed?

The construction of differentially private analyses relies on the careful introduction of uncertainty in the form of random noise. This section provides a simple example illustrating how a carefully-calibrated amount of random noise can be added to the outcome of an analysis in order to provide privacy protection. This explanation is somewhat technically involved, and a first-time reader may choose instead to skip ahead to Section 7.2.

> Consider computing an estimate of the number of HIV-positive individuals in a sample, where the sample contains $n = 10,000$ individuals of whom $m = 38$ are HIV-positive. In a differentially private version of the computation, random noise Y is introduced into the count so as to hide the contribution of a single individual. That is, the result of the computation would be $m' = m + Y = 38 + Y$ instead of $m = 38$.

The magnitude of the random noise Y affects both the level of privacy protection provided and the accuracy of the count. For instance, a larger amount of noise would result in better privacy protection and worse accuracy—and vice versa. The magnitude of Y depends on the privacy loss parameter ϵ, where a smaller value of ϵ is associated with a larger noise magnitude.

When choosing the noise distribution, one possibility is to sample the random noise Y from a normal distribution with zero mean and standard deviation $1/\epsilon$.[67] Because the choice of the value of ϵ is inversely related to the magnitude of the noise introduced by the analysis, the mechanism is designed to provide a quantifiable tradeoff between privacy and utility. Consider the following example.

> A researcher uses the estimate m', as defined in the previous example, to approximate the fraction p of HIV-positive people in the population. The computation would result in the estimate
>
> $$p' = \frac{m'}{n} = \frac{38 + Y}{10,000}.$$
>
> For instance, suppose the sampled noise is $Y = 4.2$. Then, the estimate would be
>
> $$p' = \frac{38 + Y}{10,000} = \frac{38 + 4.2}{10,000} = \frac{42.2}{10,000} = 0.42\%,$$
>
> whereas without added noise, the estimate would have been $p = 0.38\%$.

[67] More accurately, the noise Y is sampled from the Laplace distribution with zero mean and standard deviation $\sqrt{2}/\epsilon$. The exact shape of the noise distribution is important for proving that outputting $m + Y$ preserves differential privacy, but can be ignored for the current discussion.

7.1 Two sources of error: sampling error and added noise

Note that there are two sources of error in estimating p: sampling error and added noise. The first source, sampling error, would cause m to differ from the expected $p \cdot n$ by an amount of roughly

$$|m - p \cdot n| \approx \sqrt{p \cdot n}.$$

For instance, consider how the researcher from the example above would calculate the sampling error associated with her estimate.

> The researcher reasons that m' is expected to differ from $p \cdot 10,000$ by roughly
>
> $$\sqrt{p \cdot 10,000} \approx \sqrt{38} \approx 6.$$
>
> Hence, the estimate 0.38% is expected to differ from the true p by approximately
>
> $$\frac{6}{10,000} = 0.06\%,$$
>
> even prior to the addition of the noise Y by the differentially private mechanism.

The second source of error is the addition of random noise Y in order to achieve differential privacy. This noise would cause m' and m to differ by an amount of roughly

$$|m' - m| \approx 1/\epsilon.$$

The researcher in the example would calculate this error as follows.

> The researcher reasons that, with a choice of $\epsilon = 0.1$, she should expect $|m' - m| \approx 1/0.1 = 10$, which can shift p' from the true p by an additional $\frac{10}{10,000} = 0.1\%$.
>
> Taking both sources of noise intro account, the researcher calculates that the difference between noisy estimate p' and the true p is roughly
>
> $$0.06\% + 0.1\% = 0.16\%.$$
>
> Because the two sources of noise are statistically independent, the researcher can use the fact that their variances add to produce a slightly better bound:
>
> $$|p' = p| \approx \sqrt{0.06^2 + 0.1^2} = 0.12\%.$$

Generalizing from this example, we find that the standard deviation of the estimate p' (hence the expected difference between p' and p) is of magnitude roughly

$$|p' - p| \approx \sqrt{p/n} + 1/n\epsilon,$$

which means that for a large enough sample size n the sampling error would far exceed the noise added for the purposes of privacy protection.

Note also that the literature on differentially private algorithms has identified many other noise introduction techniques that result in better accuracy guarantees than the simple technique used in the examples above. Such techniques are especially important for more complex analyses, for which the simple noise addition technique discussed in this section is often sub-optimal in terms of accuracy.

7.2 What types of analyses can be performed with differential privacy?

A large number of analyses can be performed with differential privacy guarantees. The following is a non-exhaustive list of analyses for which differentially private algorithms are known to exist:

- **Count queries:** The most basic statistical tool, a count query returns an estimate of the number of individual records in the data satisfying a specific predicate. For example, a count query could be used to return the number of records corresponding to HIV-positive individuals in a sample. Differentially private answers to count queries can be obtained through the addition of random noise, as demonstrated in the detailed example found above in Section 7.

- **Histograms:** A histogram contains the counts of data points as they are classified into disjoint categories. For example, in the case of numerical data, a histogram shows how data are classified within a series of consecutive non-overlapping intervals. A **contingency table (or cross tabulation)** is a special form of a histogram representing the interrelation between two or more variables. The categories of a contingency table are defined as conjunctions of attribute variables. Differentially private histograms and contingency tables provide noisy counts for the data classified in each category.

- **Cumulative distribution function (CDF):** For data over an ordered domain, such as age (where the domain is integers, say, in the range $0-100$), or annual income (where the domain is real numbers, say, in the range $\$0.00 - \$1,000,000.00$), a cumulative distribution function depicts for every domain value x an estimate of the number of data points with a value up to x. A CDF can be used for computing the median of the data points (the value x for which half the data points have value up to x) and the interquartile range, among other statistics. A differentially private estimate of the CDF introduces noise that needs to be taken into account when the median or interquartile range is computed from the estimated CDF.[68]

- **Linear regression:** Social scientists are often interested in modeling how some dependent variable varies as a function of one or more explanatory variables. For instance, a researcher may seek to understand how a person's health depends on her education and income. In linear regression, an underlying linear model is assumed, and the goal of the computation is to fit a linear model to the data that minimizes a measure of "risk" (or "cost"), usually the sum of squared errors. Using linear regression, social scientists can learn to what extent a linear model explains their data, and which of the explanatory variables correlates best with the dependent variable. Differentially private implementations of linear regression introduce noise in its computation. Note that, while this noise may in some cases hide existing correlations

[68] For a more in depth discussion of differential privacy and CDFs, see Daniel Muise and Kobbi Nissim, "Differential Privacy in CDFs," Slide Deck (2016), http://privacytools.seas.harvard.edu/files/dpcdf_user_manual_aug_2016.pdf.

in the data, researchers are engaged in ongoing work towards the development of algorithms where this undesirable effect of noise addition can be controlled and minimized.

- **Clustering:** Clustering is a data analysis technique which involves grouping data points into clusters, such that points in the same cluster are more similar to each other than to points in other clusters. Data scientists often use clustering as an exploratory tool to gain insight into their data and identify the data's important sub-classes. Researchers are developing a variety of differentially private clustering algorithms, and such tools are likely to be included in future privacy-preserving tool kits for social scientists.

- **Classification:** Classification is the problem of identifying which of a set of categories a data point belongs in, based on a training set of examples for which category membership is known. Data scientists often utilize data samples that are pre-classified (e.g., by experts) to train a classifier which can later be used for labeling newly-acquired data samples. Theoretical work has shown that it is possible to construct differentially private classification algorithms for a large collection of classification tasks, and, furthermore, that, at least in principle, the performance of these classification algorithms is comparable with the performance of similar non privacy preserving algorithms.

8 Practical challenges to using differential privacy

In this section, we discuss some of the practical challenges to using differential privacy, including challenges related to the accuracy of differentially private statistics, and challenges due to the degradation of privacy that results from multiple analyses. It is important to note that the challenges of producing accurate statistics while protecting privacy and addressing composition are not unique to differential privacy. It is a fundamental law of information that privacy risk grows with the use of data, and hence this risk applies to any disclosure control technique. Traditional statistical disclosure limitation techniques, such as suppression, aggregation, and generalization, often reduce accuracy and are vulnerable to loss in privacy due to composition, and the impression that these techniques do not suffer accumulated degredation in privacy is merely due to the fact that these techniques have not been analyzed with the higher level of rigor that differential privacy is.[69] A rigorous analysis of the effect of composition is important for establishing a robust and realistic understanding of how multiple statistical computations affect privacy.

8.1 Accuracy

Differentially private computations rely on the introduction of random noise that is sufficiently large to hide the contribution of roughly any subset of (roughly) $1/\epsilon$ individuals. As a consequence, differentially private computations are less accurate than the statistics one could directly compute on the data. Put differently, differential privacy increases the minimal sample size required to produce accurate statistics.

Much of the ongoing research on differential privacy is focused on understanding and improving this tradeoff, i.e., how to obtain the maximum possible utility from data while preserving differential

[69] For a discussion of privacy and utility with respect to traditional statistical disclosure limitation techniques, see Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, and Ashwin Machanavajjhala, Privacy-Preserving Data Publishing, Foundations and Trends in Databases 2.1-2 (2009): 1-167.

privacy.[70] In practice, the amount of noise that is added to differentially private analyses makes it difficult to obtain much utility from small- to moderately-sized datasets. As a rule of thumb, almost no utility is expected from datasets containing $1/\epsilon$ or fewer records.[71]

For certain types of analyses, procedures have been developed for estimating the accuracy of an analysis based on properties of the collected data. These procedures take as input the number of records, a value for ϵ, and the ranges of numerical and categorical fields, among other parameters, and produce bounds on the accuracy for a variety of statistical computations. Alternatively, a desired accuracy may be given as input instead of ϵ, and the computation results in a value for ϵ that would provide this level of accuracy. Figure 4 illustrates an example of a cumulative distribution function and the results of its (noisy) approximation with different settings of the privacy parameter ϵ.[72] Procedures for estimating the accuracy of an analysis are being developed for practical implementations of differential privacy, including the tools that are being developed for Harvard's Dataverse project, as discussed below.

8.2 The "privacy budget"

One can think of the parameter ϵ as determining the overall privacy protection provided by a differentially private analysis. Intuitively, ϵ determines "how much" of an individual's privacy an analysis may utilize, or, alternatively, by how much the risk to an individual's privacy can increase. A smaller value for ϵ implies better protection, i.e., less risk to privacy. Conversely, a larger value for ϵ implies worse protection, i.e., higher potential risk to privacy. In particular, $\epsilon = 0$ implies perfect privacy, i.e., the analysis does not increase any individual's privacy risk at all. Unfortunately, analyses that satisfy differential privacy with $\epsilon = 0$ must completely ignore their input data and therefore are useless.

We can also think of ϵ as a "privacy budget" to be spent by analyses of individuals' data. If a single analysis is expected to be performed on a given set of data, then one might allow this analysis to exhaust the entire privacy budget ϵ. However, a more typical scenario is that several analyses are expected to be run on a dataset, and hence one needs to calculate the total utilization of the privacy budget by these analyses.

Fortunately, a number of composition theorems have been developed for differential privacy, as mentioned above in Section 3.2. In particular, these theorems state that the composition of two differentially private analyses results in a privacy loss that is bounded by the sum of the privacy losses of each of the analyses.

To understand how overall privacy loss is accounted for in this framework, consider the following example.

---

[70] We use the term *accuracy* somewhat informally to refer to the quality of information that is produced by an analysis. Introduction of random noise often results in a reduction in accuracy and hence in the quality of the information produced. Note that what accuracy means, and how accuracy is measured, differs across various analyses and applications. For example, a researcher interested in estimating the average income of a given population may care about the absolute error of this estimate, i.e., the difference between the real average and the estimate, whereas a researcher interested in the median income may care about the difference between the number of respondents whose income is below the estimate and the number of respondents whose income is above the estimate.

[71] An exception is when the amplification technique known as "secrecy of the sample" is used. See Section 12 for a discussion on this topic.

[72] This figure first appeared in Daniel Muise and Kobbi Nissim, "Differential Privacy in CDFs," Slide Deck (2016), http://privacytools.seas.harvard.edu/files/dpcdf_user_manual_aug_2016.pdf.

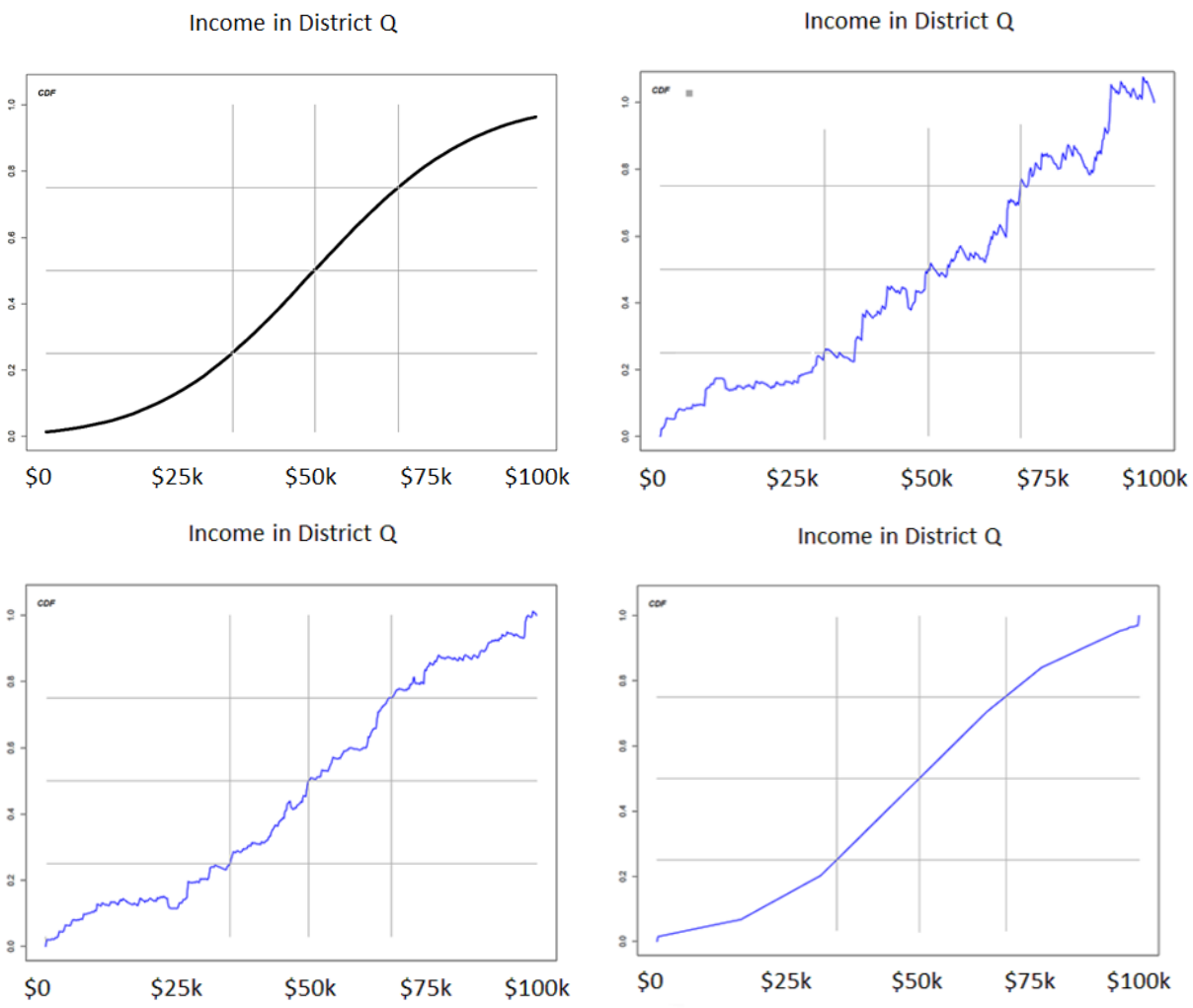

Figure 4: An example of the outcome of a differentially private computation of the cumulative distribution function (CDF) of income in District Q. The top left graph presents the original CDF (without noise) and the subsequent graphs show the result of applying differentially private computations of the CDF with ϵ values of 0.005 (top right), 0.01 (bottom left), and 0.1 (bottom right). Notice that, as smaller values of ϵ imply better privacy protection, they also imply less accuracy due to noise addition compared to larger values of ϵ.

Suppose a data analyst using a differentially private analysis tool is required to do so while maintaining differential privacy with an overall privacy loss parameter $\epsilon = 0.1$. This requirement for the overall privacy loss parameter may be guided by an interpretation of a regulatory standard, institutional policy, or best practice, among other possibilities. It means that all of the analyst's analyses, taken together, must have an epsilon value of at most 0.1.

Consider how this requirement would play out within the following scenarios:

One-query scenario: The data analyst performs a differentially private analysis with a privacy loss parameter $\epsilon_1 = 0.1$. In this case, the analyst would not be able to perform a second analysis over the data without risking a breach of the policy limiting the overall privacy loss to $\epsilon = 0.1$.

Multiple-query scenario: The data analyst first performs a differentially private analysis with $\epsilon_1 = 0.01$, which falls below the limit of $\epsilon = 0.1$. This means that the analyst can also apply a second differentially private analysis, say with $\epsilon_2 = 0.02$. After the second analysis, the overall privacy loss amounts to

$$\epsilon_1 + \epsilon_2 = 0.01 + 0.02 = 0.03,$$

which is still less than $\epsilon = 0.1$, and hence allows the analyst to perform additional analyses before exhausting the budget.

The multiple-query scenario can be thought of as if the data analyst has a *privacy budget* of $\epsilon = 0.1$ that is consumed incrementally as she performs differentially private analyses, until the budget has been exhausted. Performing additional analyses after the overall budget has been exhausted may result in a privacy parameter that is larger (i.e., worse) than ϵ. Any further use would result in a privacy risk that is too significant.

Note that, in the sample calculation for the multiple-query example, we bounded the accumulated privacy risk simply by adding the privacy parameters of each analysis. It is in fact possible to obtain better bounds on the accumulation of the privacy loss parameter than suggested by this example. Various tools for calculating the bounds on the accumulated privacy risks in real-world settings using more sophisticated approaches are currently under development.

9 Tools for differentially private analysis

At the time of this writing, differential privacy is transitioning from a purely theoretical mathematical concept to one that underlies software tools for practical use by analysts of privacy-sensitive data. This section briefly reviews some of these newly-emerging tools, with a particular focus on the tools that inspired the drafting of this document.

9.1 Differential privacy in Harvard's Dataverse project

The *Privacy Tools for Sharing Research Data* project[73] at Harvard University develops tools to help social scientists and other researchers collect, analyze, and share data while providing privacy protection for individual research subjects. To this end, the project seeks to incorporate definitions and algorithmic tools from differential privacy into Dataverse, an open-source software application developed at Harvard. Dataverse provides a preservation and archival infrastructure that enables institutions to host data repositories through which researchers can upload their data or access data made available by other researchers for the purposes of replication or secondary research.

[73] Harvard Privacy Tools Project, `http://privacytools.seas.harvard.edu`.

New privacy tools being developed for integration with the Dataverse infrastructure include a private data sharing interface, *PSI* [7], which facilitates data exploration and analysis using differential privacy. The PSI interface provides guidance for users, who are not necessarily privacy experts, on how to partition a limited privacy budget among the many statistics to be produced or analyses to be run, as well as on how to interpret the noisy results produced by a differentially private algorithm. PSI also offers a basic collection of tools for producing differentially private statistics whose results can be visualized using *TwoRavens*,[74] a browser-based interface for exploring and analyzing data. Through the differentially private access enabled by PSI, researchers will be able to perform rough preliminary analyses of privacy-sensitive datasets that currently cannot be safely shared. Such access will help researchers determine whether it is worth the effort to apply for full access to the raw data.

PSI is also being designed to integrate with other tools available through Dataverse, such as *DataTags*,[75] which are simple, iconic labels that categorically describe certain requirements for handling privacy-sensitive data. Each DataTag maps to a different set of transfer, storage, and access requirements, from completely open data (a "blue" tag) to maximally-protected data (a "crimson" tag) [21]. When a researcher initiates a deposit of a dataset containing personal information into Dataverse, she may proceed through a manual or automated process for assigning a DataTag to the dataset based on legal and institutional requirements. A DataTag can also be assigned outside of Dataverse, e.g., by the data owner with the aid of an automated decision support tool or by an expert based on direct examination of the dataset. From the time of assignment, the Dataverse repository will ensure that the storage and access requirements specified by the DataTag are met. A dataset's DataTag will also be made available via the Dataverse API, so that it can be accessed by various data management and analysis tools including PSI. The Privacy Tools project seeks to develop tools using the DataTags framework to denote handling policies for different versions of a dataset or for statistics derived from a dataset. For example, while a raw version of a privacy-sensitive dataset might be assigned a more restrictive DataTag (e.g., "red" or "crimson") that enables access only by approved researchers, differentially private statistics derived from the data might be assigned a less restrictive DataTag (e.g., "green") that enables access by any user registered with the Dataverse repository. In addition, members of the Privacy Tools project are assessing the privacy protection guaranteed by different settings of the differential privacy parameters (ϵ and δ), so that they can make recommendations regarding the values of these parameters that are appropriate for producing statistics from a dataset that has been assigned a given DataTag.

9.2 Other experimental implementations of differential privacy

Several other experimental systems enable data analysts to construct privacy-preserving analyses without requiring an understanding of the subtle technicalities of differential privacy. Systems such as Privacy Integrated Queries (PINQ) [15], Airavat [20], and GUPT [17] aim to make it easier for users to write programs that are guaranteed to be differentially private, either by composition of differentially private building blocks [15, 8], or through the use of general frameworks such as "partition-and-aggregate" or "subsample-and-aggregate" [18] to convert non-private programs into differentially private ones [20, 17]. These systems rely on a common approach: they keep the data

[74] TwoRavens, `http://datascience.iq.harvard.edu/about-tworavens`.
[75] DataTags, `http://datatags.org`.

safely stored and allow users to access them only via a programming interface which guarantees differential privacy. They also afford generality, enabling one to design many types of differentially private programs that are suitable for a wide range of purposes. However, note that most of these systems do not provide much guidance for a lay user who has limited expertise in programming. Moreover, they do not provide much guidance on deciding how to partition a limited privacy budget among many statistics or analyses, or how to interpret the noisy results given by a differentially private algorithm.

9.3 Tools for specific data releases or specific algorithms

There have been a number of successful applications of differential privacy with respect to specific, structured sources of data, including commuter patterns [14], mobility data [16], client-side software data [6], genome-wide association studies [12], location histhory data [5], and usage patterns for phone technology [11]. In these settings, differential privacy experts carefully optimize the choice of differentially private algorithms and the partitioning of the privacy budget to maximize utility for the particular data source. These tools are specific to the type of data they are designed to handle, and they cannot be applied in contexts in which the collection of data sources and the structure of the datasets are too heterogenous to allow for such optimizations.

Beyond these examples, there is a vast literature on the design of differentially private algorithms for specific data analysis tasks, including substantial experimental work on comparing and optimizing such algorithms across a wide range of datasets. For example, the recent work on DP-Bench [9], a framework for standardized evaluation of the accuracy of privacy algorithms, provides a thorough comparison of different algorithms and different ways of optimizing them.[76]

[76] See also DPComp, `https://www.dpcomp.org`.

Part IV
Advanced topics

We conclude with some advanced topics for readers interested in exploring differential privacy further. This section explains the significance of differential privacy being a property of a computation rather than a property of the result of a computation, discusses the protection differential privacy can provide for small groups of individuals, and introduces the concept of the secrecy of the sample.

10 Differential privacy: A property of the analysis (not its specific outcome)

Many disclosure limitation techniques restrict the outcome of a computation rather than restrict the computation itself. For example, the data anonymization technique k-anonymity requires that tabular data be transformed such that the identifying attributes that appear for each person in the k-anonymized data release are identical to that of at least $k-1$ other individuals in the data. Therefore, k-anonymity is defined as a property of the anonymized data output, and it imposes no further restrictions on the process used to create a k-anonymized data output. Note, however, that many possible k-anonymized outputs exist for a given dataset. A hypothetical data processor applying k-anonymity could, either maliciously or unwittingly, choose among the possible k-anonymous outputs in a way that is dependent on a sensitive attribute about an individual in the data. For example, if for a given dataset there exist two possible k-anonymized outputs T_1 and T_2, the processor may decide to output T_1 if John is HIV-positive and T_2 otherwise, thus compromising John's privacy. While we do not claim real implementations of k-anonymity suffer from this problem, the notion of k-anonymity does not preclude it.

The requirement of differential privacy is of a different nature. Rather than restricting the outcome of a differentially private computation, the definition restricts the process used to produce the computation. To understand what we mean by this, consider what happens when a statistical analysis is performed over privacy-sensitive data. Recall that, in order to yield any information of interest, the outcome of an analysis must depend on the input data. As a result, the outcome necessarily exhibits some non-zero leakage of information about the input data. The privacy concern is that an individual or organization observing the outcome of this computation would use it to infer personal information that is specific to an individual. We will consider a few illustrative examples to understand how such a privacy breach can occur.

> A collection of medical records from State Hospital includes John's medical records, which describe treatment related to an HIV diagnosis. A computation is performed on these records and outputs the following line:
>
> $$John, HIV+$$

Is John's privacy breached as a result of this computation, in the sense that it has revealed some personal information about John? The answer is *not necessarily*. For example, suppose this computation ignores its input data altogether and always outputs "John, HIV+." In this case, there

is no functional dependence between the HIV status in John's medical records and the outcome of the computation. Therefore, the mechanism does not leak any information about John.

In another extreme example, we can also see that omitting John from the outcome is not sufficient to guarantee privacy protection for John.

> A privacy-preserving mechanism transforms this collection of medical records from State Hospital by redacting all medical records pertaining to HIV-positive patients. As a result, John's records are redacted from the medical records included in the output.

It may be tempting to assume that, because John's medical records were omitted from the output, his privacy has been protected. However, the mere fact that John's information was redacted can result in a breach of his privacy. Consider the following example.

> Eve knows that John was a patient at State Hospital. Eve reviews the records that State Hospital has made available to researchers, knowing that they have been redacted of records from HIV-positive patients. Eve, noticing that John's record is absent from the redacted records, concludes that John is HIV-positive.

These examples illustrate that it is the *functional relationship between a computation's input and output* that determines to what extent personal information about an individual can be learned from the output of the computation. This intuition holds even in more complex settings, such as mechanisms in which the relationship between the input data and outcome is randomized.

The definition of differential privacy follows this intuition closely. Differential privacy is not a property of a specific outcome; rather, it is a property that a computation does or does not have. To satisfy differential privacy, the behavior of an analysis should not change noticeably when John's (or any other single individual's) information is added to or removed from the input.

11 Group privacy

By holding individuals' opt-out scenarios as the relevant baseline, the definition of differential privacy directly addresses disclosures of information localized to a single individual. However, in many cases, information may be shared between multiple individuals. For example, relatives may share an address or certain genetic attributes.

How does differential privacy protect information of this nature? Consider the opt-out scenario for a group of k individuals. This is the scenario in which the personal information of all k individuals is omitted from the input to the analysis. For instance, John and Gertrude's opt-out scenario is the scenario in which both John's and Gertrude's information is omitted from the input to the analysis.

Recall that the parameter ϵ controls by how much the real-world scenario can differ from any individual's opt-out scenario. It can be shown that the difference between the real-world and opt-out scenarios of a group of k individuals grows to at most

$$k \cdot \epsilon.$$

This means that the privacy guarantee degrades moderately as the size of the group increases. Effectively, a meaningful privacy guarantee can be provided to groups of individuals of a size of up to about

$$k \approx 1/\epsilon$$

individuals. However, almost no protection is guaranteed to groups of

$$k \approx 10/\epsilon$$

individuals or greater. This is the result of a design choice to not *a priori* prevent analysts using differentially private mechanisms from discovering trends across moderately-sized groups.

12 Amplifying privacy: Secrecy of the sample

As discussed in Section 8, differential privacy limits accuracy, and the extent of the inaccuracy depends inversely on the privacy parameter ϵ. Sometimes, the dataset used as input to a differentially private mechanism is a random sample from a large population, as in the following example.

> Alice, a researcher at State University, collected personal information from individuals in a study exploring the relationship between coffee consumption, economic status, and health status. The personal information she collected in this study is based on a uniformly random and secret sample of $3,000$ individuals living in the city of Boston.

Because Alice's study uses a uniformly random sample,[77] and, furthermore, the identities of the participating individuals are kept confidential, Alice can apply a theorem in differential privacy known as "secrecy of the sample." This theorem effectively allows for a savings in the privacy parameter ϵ that corresponds to the ratio of sizes between the dataset and the larger population. For instance, for a population the size of the city of Boston, approximately $600,000$, the savings in ϵ can be $3,000/600,000 = 0.05$. This means that greater accuracy, corresponding to a 0.05 decrease in epsilon, can be provided for the differentially private analyses performed on the data from Alice's study.

This topic comes with two notes of caution. First, sampling from the sampling frame is usually not uniform in practice. Alice should therefore be conservative in her estimate of the underlying population. For example, if Alice draws her survey sample from a Boston phonebook, then she should take the underlying population size to be no larger than the number of Boston residents who are listed in the phonebook. Second, the effective decrease in ϵ is conditioned on the identities of the sampled individuals being kept secret. This may be a hard condition to maintain in practice. For example, if Alice sends surveyors to respondents' homes, then their neighbors may learn that they participated in Alice's research study. A more subtle consideration is that secrecy of the sample also requires the identities of individuals who have *not* been sampled to be kept secret.

[77] By *uniformly random* we mean that each individual in the sampling frame is selected to be in the sample with equal probability and independently of the other individuals in the sampling frame.

Part V
Summary

Differential privacy provides a formal, quantifiable measure of privacy. It has been established by a rich and rapidly evolving theory that enables one to reason with mathematical rigor about privacy risk. Quantification of privacy is achieved by the privacy loss parameter ϵ, which controls, simultaneously for every individual contributing to the analysis, the deviation between one's opt-out scenario and the actual execution of the differentially private analysis. This deviation can grow as an individual participates in additional analyses, but the overall deviation can be bounded as a function of ϵ and the number of analyses performed. This amenability to *composition* is a unique feature of differential privacy. While it is not the only framework that quantifies a notion of risk for a single analysis, it is currently the only framework with quantifiable guarantees on the risk resulting from a composition of several analyses.

In other words, the parameter ϵ can be interpreted as bounding the excess risk to an individual resulting from her data being used in analysis (compared to her risk when her data are not being used). Indirectly, the parameter ϵ also controls the accuracy to which a differentially private computation can be performed. For researchers making privacy-sensitive data available through a differentially private tool, the interface of the tool may allow them to choose to produce a variety of differentially private summary statistics while maintaining a desired level of privacy (quantified by an accumulated privacy loss parameter), and then compute summary statistics with formal privacy guarantees.

Differential privacy can be used to make more data available in a privacy-preserving way, which can help researchers showcase their data to other researchers who may be interested in using the data in their own studies. This, in turn, can further the progress of scientific discovery and build the reputations of the researchers collecting and sharing data. For researchers seeking data for their own studies, differentially private summary statistics could provide a basis for determining whether a particular dataset is likely to be useful to them—and hence whether they should proceed with a negotiation for obtaining the data. In this way, differentially private tools hold promise for opening up access to data that cannot currently be shared, thereby enabling new analyses to be performed and ultimately advancing the state of scientific knowledge.

Further reading

Differential privacy was introduced in 2006 by Dwork, McSherry, Nissim and Smith [3]. This document's presentation of the opt-out scenario vs. real-world computation is influenced by [1], and its risk analysis is influenced by [13]. For other presentations of differential privacy, see [2] and [10]. For a thorough technical introduction to differential privacy, see [4].

References

[1] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006)*, pages 1–12, 2006. http://dx.doi.org/10.1007/11787006_1.

[2] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011. http://doi.acm.org/10.1145/1866739.1866758.

[3] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In 3^{rd} *Theory of Cryptography Conference*, pages 265–284, 2006. http://dx.doi.org/10.1007/11681878_14.

[4] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. http://dx.doi.org/10.1561/0400000042.

[5] Andrew Eland. Tackling urban mobility with technology. Google Policy Europe Blog, 2015. https://europe.googleblog.com/2015/11/tackling-urban-mobility-with-technology.html.

[6] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In Gail-Joon Ahn, Moti Yung, and Ninghui Li, editors, *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067. ACM, 2014. http://doi.acm.org/10.1145/2660267.2660348.

[7] Marco Gaboardi, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil P. Vadhan. PSI (Ψ): a Private data Sharing Interface. *CoRR*, abs/1609.04340, 2016. http://arxiv.org/abs/1609.04340.

[8] Andreas Haeberlen, Benjamin C. Pierce, and Arjun Narayan. Differential privacy under fire. In *Proceedings of the 20th USENIX Security Symposium*, August 2011. http://www.cis.upenn.edu/~ahae/papers/fuzz-sec2011.pdf.

[9] Michael Hay, Ashwin Machanavajjhala, Gerome Miklau, Yan Chen, and Dan Zhang. Principled evaluation of differentially private algorithms using DPBench. In *Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16)*, 2016. http://dl.acm.org/citation.cfm?id=2882931.

[10] Ori Heffetz and Katrina Ligett. Privacy and data-based research. *Journal of Economic Perspectives*, 28(2):75–98, 2014. http://www.aeaweb.org/articles.php?doi=10.1257/jep.28.2.75.

[11] Apple Press Info. Apple previews iOS 10, the biggest iOS release ever, 2016. https://www.apple.com/pr/library/2016/06/13Apple-Previews-iOS-10-The-Biggest-iOS-Release-Ever.html.

[12] Xiaoqian Jiang, Yongan Zhao, Xiaofeng Wang, Bradley Malin, Shuang Wang, Lucila Ohno-Machado, and Haixu Tang. A community assessment of privacy preserving techniques for human genomes. *BMC Medical Informatics and Decision Making*, 14(Suppl 1)(S1), 2014. https://www.ncbi.nlm.nih.gov/pubmed/25521230.

[13] Shiva Prasad Kasiviswanathan and Adam Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, abs/0803.3946, 2008. http://arxiv.org/abs/0803.3946.

[14] Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *Proceedings of the 24th International Conference on Data Engineering (ICDE 2008)*, pages 277–286, 2008. http://dx.doi.org/10.1109/ICDE.2008.4497436.

[15] Frank McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 International Conference on Management of Data (SIGMOD '09)*, pages 19–30, 2009. http://doi.acm.org/10.1145/1559845.1559850.

[16] Darakhshan J. Mir, Sibren Isaacman, Ramón Cáceres, Margaret Martonosi, and Rebecca N. Wright. DP-WHERE: differentially private modeling of human mobility. In *Proceedings of the 2013 IEEE International Conference on Big Data*, pages 580–588, 2013. http://dx.doi.org/10.1109/BigData.2013.6691626.

[17] Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David E. Culler. GUPT: privacy preserving data analysis made easy. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '12)*, pages 349–360, 2012. http://doi.acm.org/10.1145/2213836.2213876.

[18] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the* 39^{th} *ACM Symposium on Theory of Computing (STOC '07)*, 2007. http://doi.acm.org/10.1145/1250790.1250803.

[19] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701, 2010. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006.

[20] Indrajit Roy, Srinath T. V. Setty, Ann Kilzer, Vitaly Shmatikov, and Emmett Witchel. Airavat: Security and privacy for MapReduce. In *Proceedings of the 7th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2010)*, pages 297–312, 2010. http://www.usenix.org/events/nsdi10/tech/full_papers/roy.pdf.

[21] Latanya Sweeney, Merce Crosas, and Michael Bar-Sinai. Sharing sensitive data with confidence: The Datatags system. *Technology Science*, 2015. http://techscience.org/a/2015101601/.