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

(Preliminary version)

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

[1]Department of Computer Science, Georgetown University.
kobbi.nissim@georgetown.edu.
[2]Center for Research on Computation and Society, Harvard University.
{tsteinke|mbun|salil}@seas.harvard.edu.
[3]Berkman Klein Center for Internet & Society, Harvard University.
{awood|dobrien}@cyber.law.harvard.edu.
[4]University at Buffalo, The State University of New York. gaboardi@buffalo.edu.

March 3, 2017

Abstract

This document is a primer on *differential privacy*, which is a 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. Using intuitive illustrations and limited mathematical formalism, this document 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 scientists 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.

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 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.

Contents

Audience

This document is written for a non-technical audience. In particular, it is intended to serve as a resource for social science researchers, such as sociologists, psychologists, behavioral economists, and political scientists, who collect and analyze privacy-sensitive personal data and make decisions whether and how to share their research data and results with others. It is also intended to be helpful more broadly to legal scholars and policymakers, as they consider how current and future legal frameworks and instruments will apply to tools based on formal privacy models such as differential privacy.

The goal for this document is to introduce the reader to the concept of *differential privacy*, a new formal mathematical model of privacy protection. Differential privacy underlies some of the 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]

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, so that they can make 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.

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,

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

[2] See Section 8.2 below for a list of other implementations of differential privacy.

and regulatory agencies. Notably for researchers in the United States, research involving human 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

[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.

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.

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. Weaknesses have also been found with respect to other approaches to privacy. 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.

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. 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.

2 What is the differential privacy guarantee?

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 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.

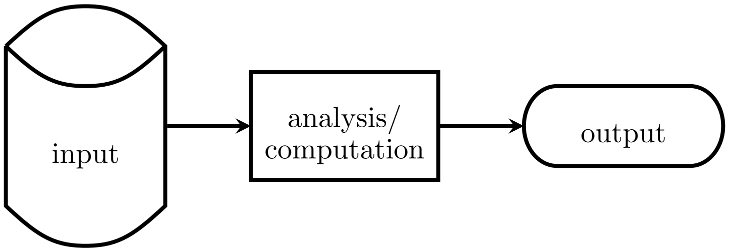

Figure 1: An analysis (or computation) transforms input data into some output.

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.[4]

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

> 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

[4] See Section 10.1 for further discussion of the importance of defining privacy as a property of the computation.

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.

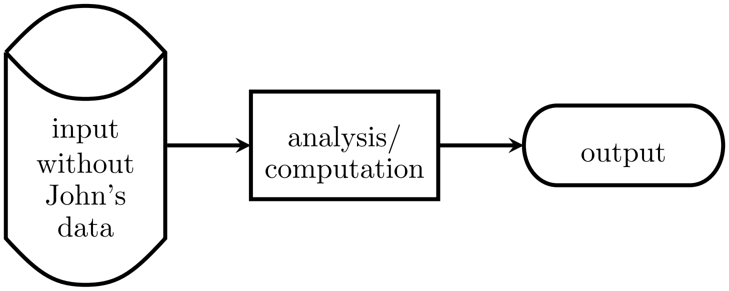

Figure 2: John's opt-out scenario.

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. 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.

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 hence 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.[5]

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.

> 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).

[5] 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.

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 7.2 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, 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 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 6 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.[6] 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.

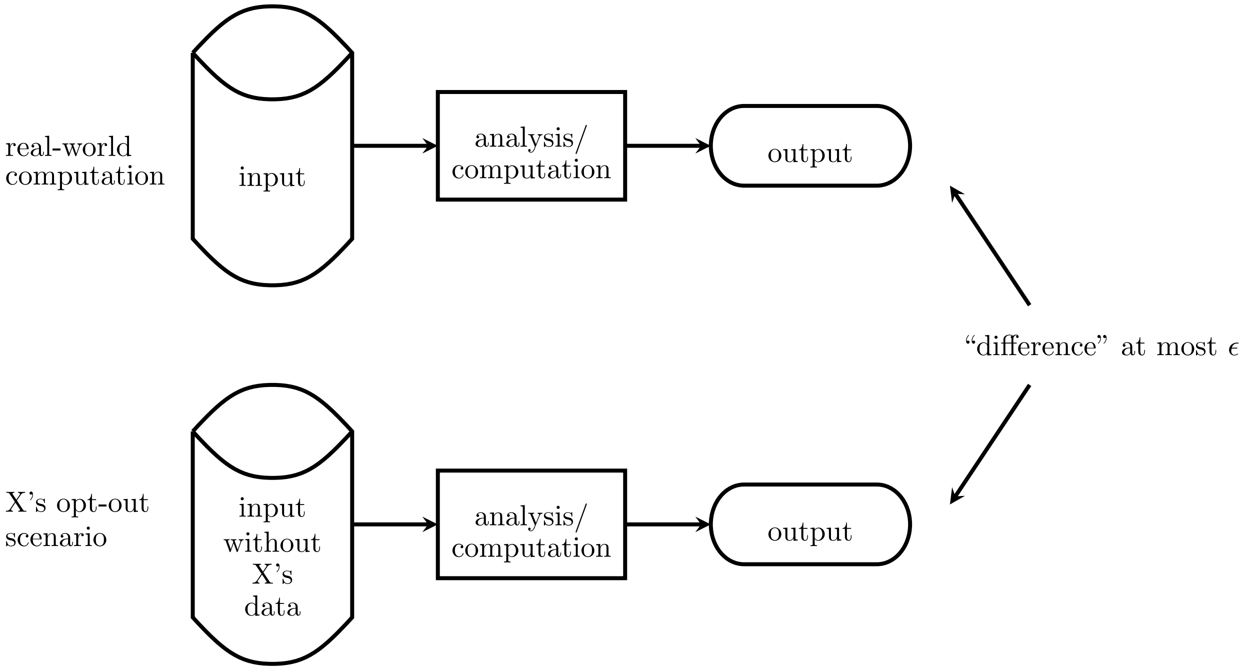

Figure 3: Differential privacy. The maximum deviation between the opt-out scenario and real-world computation should hold simultaneously for each individual X whose information is included in the input.

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.

[6] 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.

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.[7]

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.[8] An *event* in this case is simply a subset of the potential answers for the analysis. For example, we can define the following event:

$$E_1: \quad \text{the outcome of the analysis is between 0.1 and 0.2.}$$

Consider an analysis on some input data, for which event E_1 would occur with some probability p. For an analysis on the input data excluding John's data, event E_1 may occur with probability p'. The guarantee of differential privacy is that these two probabilities, p and p', are almost the same. More precisely,

$$p \leq (1+\epsilon) \cdot p' = 1.01 \cdot p'$$

and, similarly,

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

[7] For an explanation of how differentially private analyses are constructed, see Section 6 below.

[8] 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 fair die is tossed, the possible outcomes, making up the sample space, are $\{1, 2, 3, 4, 5, 6\}$. 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 $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$.

In other words, if $p' = 0.1$, then we find that p is between $0.1/1.01 \approx 0.099$ and $0.1 \cdot 1.01 = 0.101$. (Note: In this example, we assign $\epsilon = 0.01$ for concreteness. Also, this analysis is approximate and accurate only for small values of ϵ.)

Differential privacy guarantees this bound on the ratios p/p' and p'/p not only for event E_1 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:

$$E_2 : \text{the outcome of the statistical analysis is between 0.1 and 0.2.}$$

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, 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, the privacy loss parameter ($\epsilon = 0.01$, in this example) ensures that 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.[9] 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 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.

[9] 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.

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.[10]

> 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.

[10] 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`.

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 potential increase in her insurance premium, is at most $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 an informed decision based on the worst-case cost from 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%.

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.

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.

> 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 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.

5 Differential privacy and legal requirements

Social scientists and other actors 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.

Relevant legal requirements vary widely depending on the actors and institutions, types of information, and jurisdictions of the actors and research subjects involved. For example, the Federal Policy for the Protection of Human Subjects[11] requires researchers who receive US federal 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[12] and the Health Insurance Portability and Accountability Act Privacy Rule[13] 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.[14] 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[15] 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,[16] 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.

Determining whether tools for differentially private analysis can satisfy legal requirements for protecting privacy such as these is challenging for a number of reasons. Because privacy laws are

[11] 45 C.F.R. Part 46.

[12] 20 U.S.C. § 1232g; 34 C.F.R. Part 99 (2013).

[13] 45 C.F.R. Part 160 and Subparts A and E of Part 164.

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

[15] 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.

[16] 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.

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 these challenges, there are strong reasons to believe that differential privacy provides privacy protection that is consistent with many legal requirements. For instance, differential privacy provides strong privacy 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 numerous privacy laws. In addition, the use of differential privacy prevents an adversary from determining whether a given individual's personal information was included in an analysis. Differential privacy also 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. 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. 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.[17] 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 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 6.2.

> 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

[17] 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, and Salil Vadhan, Bridging the Gap between Computer Science and Legal Approaches to Privacy, Working Paper (2017).

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 $1/\epsilon$.[18] 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 $p = 0.38\%$.

6.1 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

$$\frac{6}{10,000} = 0.06\%,$$

[18] 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.

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.

6.2 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 6.

- **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.[19]

- **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 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.

[19] 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.

7 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.[20] A rigorous analysis of the effect of composition is important for establishing a robust and realistic understanding of how multiple statistical computations affect privacy.

7.1 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 privacy.[21] 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.[22]

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.[23] Procedures for estimating the accuracy of an analysis are being developed for practical

[20] 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.

[21] 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.

[22] An exception is when the amplification technique known as "secrecy of the sample" is used. See Section 10.3 for a discussion on this topic.

[23] This figure first appeared in Daniel Muise and Kobbi Nissim, "Differential Privacy in CDFs," Slide Deck (2016),

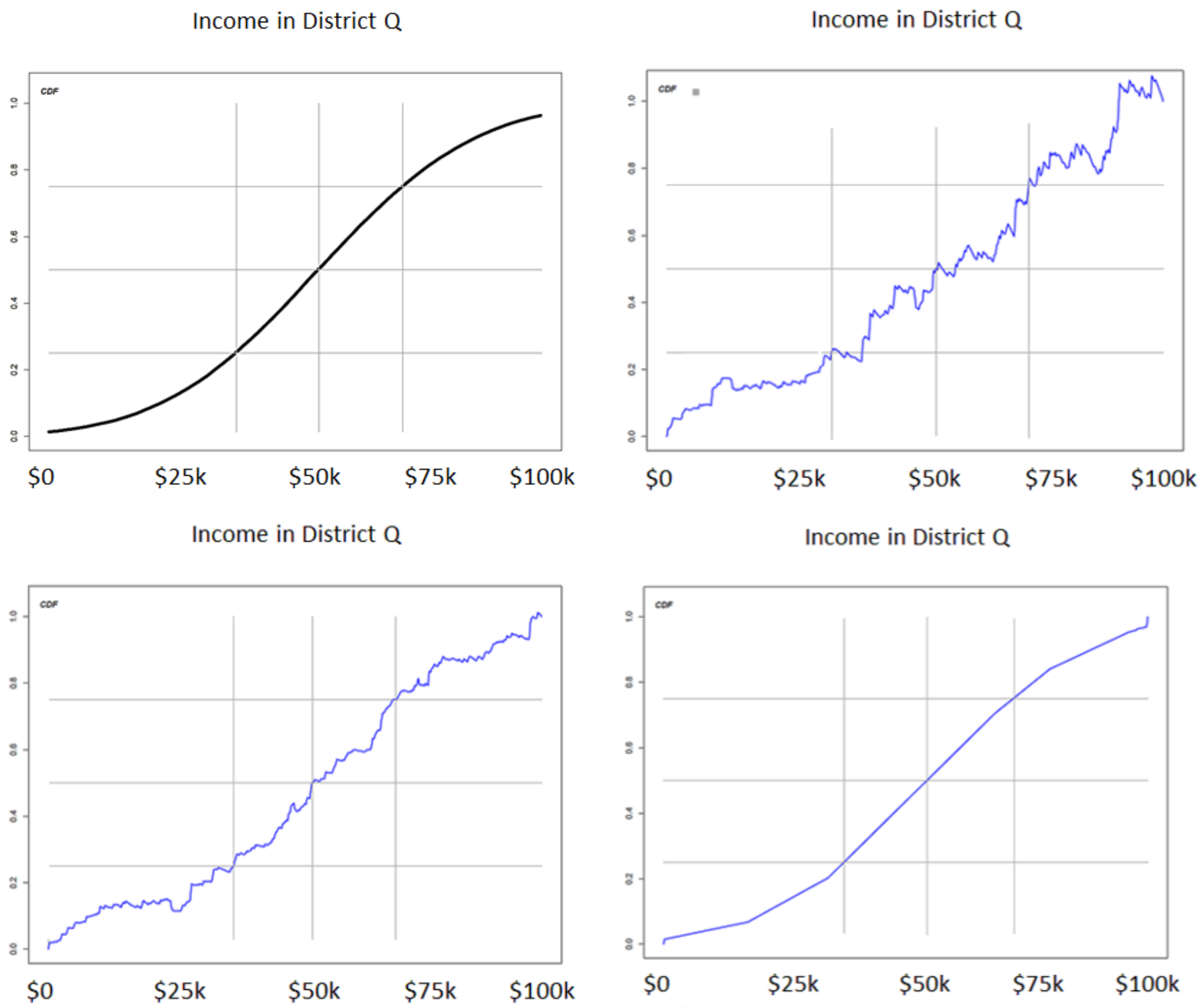

Figure 4: An example of the outcome of a differentially private computation of the cumulative distribution function (CDF) of income in District Q. The top left graph presents the original CDF (without noise) and the subsequent graphs show the result of applying differentially private computations of 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.

implementations of differential privacy, including the tools that are being developed for Harvard's Dataverse project, as discussed below.

http://privacytools.seas.harvard.edu/files/dpcdf_user_manual_aug_2016.pdf.

7.2 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.

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.

8 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.

8.1 Differential privacy in Harvard's Dataverse project

The *Privacy Tools for Sharing Research Data* project[24] 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.

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*,[25] 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*,[26] 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

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

[25] TwoRavens, `http://datascience.iq.harvard.edu/about-tworavens`.

[26] DataTags, `http://datatags.org`.

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.

8.2 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. The goal of these systems is 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 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.

8.3 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.[27]

9 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.

10 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.

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

10.1 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.

10.2 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.

10.3 Amplifying privacy: Secrecy of the sample

As discussed in Section 7, 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,[28] 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.

11 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`.

[28] 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.

[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/.