ABSTRACT

Title of dissertation: NEW NOTIONS AND MECHANISMS FOR STATISTICAL PRIVACY

Adam Groce, Doctor of Philosophy, 2014

Dissertation directed by: Professor Jonathan Katz
Department of Computer Science

Many large databases of personal information currently exist in the hands of corporations, nonprofits, and governments. The data in these databases could be used to answer any number of important questions, aiding in everything from basic research to day-to-day corporate decision-making. These questions must be answered while respecting the privacy of the individuals whose data are being used. However, even defining privacy in this setting can be difficult. The standard definition in the field is *differential privacy* [25]. During the years since its introduction, a wide variety of query algorithms have been found that can achieve meaningful utility while at the same time protecting the privacy of individuals. However, differential privacy is a very strong definition, and in some settings it can seem too strong. Given the difficulties involved in getting differentially private output to all desirable queries, many have looked for ways to weaken differential privacy without losing its meaningful privacy guarantees.

Here we discuss two such weakenings. The first is *computational differential privacy*, originally defined by Mironov et al. [56]. We find the promise of this

weakening to be limited. We show two results that severely curtail the potential for computationally private mechanisms to add any utility over those that achieve standard differential privacy when working in the standard setting with all data held by a single entity.

We then propose our own weakening, *coupled-worlds privacy*. This definition is meant to capture the cases where reasonable bounds can be placed on the adversary's certainty about the data (or, equivalently, the adversary's auxiliary information). We discuss the motivation for the definition, its relationship to other definitions in the literature, and its useful properties. Coupled-worlds privacy is actually a framework with which specific definitions can be instantiated, and we discuss a particular instantiation, *distributional differential privacy*, which we believe is of particular interest.

Having introduced this definition, we then seek new distributionally differentially private query algorithms that can release useful information without the need to add noise, as is necessary when satisfying differential privacy. We show that one can release a variety of query output with distributional differential privacy, including histograms, sums, and least-squares regression lines.

NEW NOTIONS AND MECHANISMS FOR STATISTICAL PRIVACY

by

Adam Dowlin Groce

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Jonathan Katz, Chair/Advisor
Professor Michel Cukier
Professor Hal Daume III
Professor William Gasarch
Professor Elaine Shi

Acknowledgments

I first and foremost would like to thank my advisor, Jonathan Katz, for guiding me through graduate school. Jonathan reached out to me early on and made time to help me even while on sabbatical during my first year. His advice has been invaluable, whether I was selecting problems to work on, working through the technical details of a result, or presenting the result for others. His feedback while writing papers has been especially valuable, greatly improving the quality of my technical writing. In particular, I thank him for his patience with me as my interests shifted between several different areas of research.

I would also like to thank Adam Smith for not only being an excellent collaborator, but also for a variety of professional advice and encouragement. Of course, I must also thank all the members of the Maryland Cybersecurity Center. I was fortunate to study at the University of Maryland during a time when the cryptography and security research community expanded drastically, and it has been an exciting experience.

I have had the pleasure to work with many wonderful researchers while in graduate school, and I am very grateful to everyone joined me in research, including Raef Bassily, Dov Gordon, Alex Malozemoff, Aishwarya Thiruvengadam, Arkady Yerukhimovich, and Vassilis Zikas. I have found that even the work that did not lead to a publication has been valuable and enriching. I am also grateful to Michel Cukier, Hal Daume, Bill Gasarch, and Elaine Shi for taking the time to serve on my dissertation committee.

Finally, I thank my wife Michelle, to whom this work is dedicated. The most difficult parts of graduate school are not always academic, and I could not have done it without her.

Table of Contents

Chapter 1: Introduction

Consider a hospital with a database of patient records. A medical researcher has a hypothesis regarding the cause of some disease. He suspects, say, that smoking causes lung cancer. A natural first step would be to check a database of existing patients for a correlation between lung cancer and a history of smoking. Of course, such investigations are only possible if they can be done while respecting the privacy of the patients. Currently, this privacy is usually protected through legal safeguards. The researcher must sign legal agreements promising to keep the data confidential. If the researcher makes a mistake and publishes information that violates privacy, the researcher and hospital face potential liability. The whole process involves significant bureaucratic overhead.

Not only is the red tape expensive and inconvenient, but it also prevents beneficial research from taking place. Often, the access to this sort of data is being provided largely as a favor to the researcher, and any need for costly and time-consuming oversight makes it likely that the data owner will simply avoid the project altogether. Even a benevolent data owner willing to take the time and effort to get to know researchers and draw up contracts faces practical restrictions. Data can be made accessible to a small number of people, but not to the huge research community

that is engaged when data is made freely available electronically.

Furthermore, once a researcher has access to the data, the problem of privacy inevitably reemerges later on. In most situations of this sort (certainly all academic research) the eventual goal is public dissemination of the learned information. The researcher must eventually publish something. Of course, a competent researcher will not publish raw private data, but even "summary" statistics and study results have the potential to violate the privacy of the people from whose data they are learned. The researcher must make a decision about what information is safe to publish, and in order to do this they need some sort of standard to identify privacy-violating information.

The field of statistical database privacy attempts to solve both the problem of data access and the problem of eventual publication through mathematical guarantees. Instead of allowing the researcher access to the data directly, the hospital could ask the researcher to submit the queries he would like to have run on the data. The hospital could then run those queries for him and return the results. Crucially, mathematical properties of those queries could guarantee that the results cannot be used to infer any private information about any particular patient. As a result, all information learned through these queries would be safe to publish, and the data owner would not need to rely on the decisions of the researcher to protect privacy.

In order to find queries with the relevant mathematical properties to guarantee the protection of privacy, we must first decide what mathematical properties are desired. This is a difficult question, as it requires both technical skill and an understanding of the inchoate and inconsistent idea of "privacy" that exists in the

minds of lawyers, researchers, and the general public. In fact, many early attempts at defining mathematical formulations of privacy failed to protect against all privacy violations, while others were so strong as to (unnecessarily) prevent almost any useful queries from being done.

As time progressed, researchers learned from earlier attempts, and also from the infusion of ideas from cryptography into the field. Of particular interest is *differential privacy* [25], proposed by Dwork, McSherry, Nissim, and Smith in 2006. This definition does an excellent job of capturing a reasonable understanding of privacy and converting that understanding into a workable mathematical criterion. While not without criticism (for example, [59]), it has been studied extensively, and researchers have created a large set of private queries that are known to be differentially private. We have also seen the early stages of its practical use in the real world [13, 50].

Differential privacy is a very simple definition. It is easy to work with, and it has a number of desirable properties. For example, privacy holds even if the adversary trying to make inferences about protected individuals already has partial information about them or the database in general. It is also composable, meaning that queries can be done in sequence or simultaneously by multiple users without concern.

However, in its simplicity differential privacy ignores some nuances and makes some worst-case assumptions. These provide ease of use and safety, but they also make the definition stronger than the underlying idea of "privacy" arguably requires. In order to satisfy differential privacy, a query's output must always be randomized,

generally consisting of the "true" output of some query plus a bit of random noise, generally smaller for larger databases. Many queries can be answered with low, often acceptable noise on reasonably-sized databases, but others continually resist efforts to increase accuracy. In fact, a variety of lower bounds have been shown on the amount of noise needed for specific types of queries.

For this reason, there has been interest in discarding some of the simplicity of differential privacy in favor of a weaker, more narrowly-tailored definition. It is this effort that we focus on in this thesis.

1.1 Organization and Contributions

We begin in Chapter 2 with a thorough review of the definitional work that led up to the current situation. We first discuss attacks against information releases that were thought to be private, but without any formal definitions on which to rest the claim. We then discuss early privacy definitions, including k-anonymity [70] and its intellectual descendants and the earlier work of Dalenius [16]. Understanding the motivation behind these definitions and their shortcomings is crucial to an appreciation of the questions underlying privacy definitions.

In Chapter 3 we introduce differential privacy itself. We define it, and discuss what makes it such a good definition, as well as its drawbacks and criticism it has faced. We discuss also what queries are known to be answerable under differential privacy, and lower bounds on how accurate these answers can be.

We then move on to discuss proposed relaxations of differential privacy. Chap-

ter 4 discusses *computational differential privacy*, a term for two similar (but not equivalent) definitions proposed by Mironov et al. [56] in 2009. These definitions impose minimal restrictions on the computational power of an adversary. The hope in doing so was to allow new more accurate query mechanisms. Unfortunately, we present here two impossibility results that drastically limit the potential usefulness of these definitions. While there are new mechanisms that are private only under the computational relaxation, we show that those mechanisms generally cannot produce results that are any more accurate than what can be achieved under differential privacy.

Next, we discuss another direction of relaxation, namely that of limiting the auxiliary information that an adversary might have. As discussed earlier, differential privacy protects individuals even when the adversary trying to make inferences about them already has an arbitrary amount background information. While adversaries certainly have access to some amount of this auxiliary information, it is not limitless. By not requiring privacy to hold against adversaries with unlimited auxiliary information, we can give a definition that is weaker than differential privacy, but still guarantees privacy in realistic settings. In Chapter 5 we present one such definition, *coupled-worlds (CW) privacy*.

Coupled-worlds privacy is a framework, rather than a specific definition. One can easily instantiate a particular definition with guarantees that match the understanding of privacy that is relevant to a given situation. We discuss the motivation for the framework, and contrast it with prior similarly-motivated definitions. We prove a variety of properties of the definitions instantiated using this framework,

increasing its ease of use and providing evidence that we have correctly captured the meaning of privacy.

In Chapter 6 we focus on *distributional differential privacy (DDP)*, an instantiation of CW privacy with a meaning similar to that of differential privacy. We present several DDP query mechanisms that allow for exact (deterministic) output, which is impossible under differential privacy. In particular, we show that a class of functions we call *stable* whose output can be released privately, as can truncated histograms and sums. Sums in particular are discussed at great length. For data drawn from most continuous distributions, we show that sums of d-dimensional vectors can be released privately. This is of direct interest, but also allows corollaries showing that other queries can be answered privately as well. In particular, we discuss linear regression, which we hope will be first step in the quest to show that many machine learning algorithms can be computed privately.

Chapter 2: Background

In this chapter, we begin our discussion of privacy definitions prior to our work. The next chapter will focus on differential privacy, which is central both to the field as a whole and to this thesis. However, we begin first with work that was done (mostly) before the introduction of differential privacy. This work is necessary to understand because experience with these definitions (or lack thereof) provides part of the case for many of the decisions inherent in the design of differential privacy. For example, we will attempt later to reduce the worst-case assumption differential privacy makes about available side information, but to do this safely one needs to understand why such a conservative assumption was made in the first place.

2.1 The Data Release Setting

We begin with the setting most encountered in current practice (as opposed to current academic research on the topic). This is the setting of *data release*. Here some data owner (an academic, business, or government agency, for example) has data that consists of information about a number of individuals. The owner wants to release some of this database but wants to protect the privacy of the individuals while doing so. Sometimes, of course, this could be handled by having those who see

the data sign confidentiality agreements prohibiting further distribution, but this is cumbersome and hard to enforce. It would be preferable to modify the database in some way so as to guarantee private information was protected. The central question, then, is what sort of data releases we should consider private, and how we can modify a database so that it can be released in accordance with that requirement while at the same time continuing to provide, to the greatest extent possible, the same utility as the un-modified database.

Databases here are generally thought of as tables in which a row represents the information associated with a given individual and contains a number of fields, each of which has one piece of data about that individual. The principles in question can sometimes be generalized to databases with other forms (say, graph data), but for simplicity we limit ourselves to the simpler table-like setting here.

2.2 Informal Anonymization

The most common method for protecting the privacy of individuals when databases are disclosed is *anonymization*. This is a term for the modification of a database in a way that is meant to prevent the linking of specific data with particular individuals. Historically, anonymization was the first and most widespread method of protecting private data while allowing public access. It is still used widely in practice.

Anonymization is more an art than a science, with instructions generally representing guidelines rather than complete specifications. Even when followed in

good faith, we believe these guidelines do not lead to well-protected data. As an example, consider the guidelines of the Inter-university Consortium for Political and Social Research (ICPSR), a leading public data repository run at the University of Minnesota. ICPSR offers suggested language for researchers to use when asking for private data from study participants, including the promise that "Any personal information that could identify you will be removed or changed before files are shared in any way, including with other researchers, or results are made public" [42]. ICPSR also gives suggested practices to ensure the promised privacy is maintained.

According to ICPSR, there are two kinds of potentially concerning information: direct and indirect identifiers. Direct identifiers are described [41] as follows:

> **Direct identifiers.** These are variables that point explicitly to particular individuals or units. Examples include:
>
> - Names
> - Addresses, including ZIP and other postal codes
> - Telephone numbers, including area codes
> - Social Security numbers
> - Other linkable numbers such as driver's license numbers, certification numbers, etc.
>
> *All variables directly identifying research subjects must be removed or masked prior to deposit.*

It does seem wise to remove all direct identifiers, but there are some obvious questions raised by this list. Why are ZIP codes or area codes on their own considered direct identifiers? They do not point directly to a single individual in the way that a social security number does. The real difficulty, however, comes in the discussion of indirect identifiers.

> **Indirect identifiers.** These are variables that can be problematic as they may be used together or in conjunction with other information to identify individual respondents. Examples include:

- Detailed geographic information (e.g., state, county, province, or census tract of residence)
- Organizations to which the respondent belongs
- Educational institutions (from which the respondent graduated and year of graduation)
- Detailed occupational titles
- Place where respondent grew up
- Exact dates of events (birth, death, marriage, divorce)
- Detailed income
- Offices or posts held by respondent

The first thing to note here is that the definition of "indirect identifier" is extremely broad. In principle, *any* known information about an individual could be used to help identify which row of the database is associated with that individual. Even the examples rely on a lot of ambiguous terminology. How "detailed" does an occupational title have to be in order to count as an indirect identifier? Why are non-detailed occupational titles not included? Any stipulation of occupation at all could be used to substantially narrow the number of candidate rows that could be associated with a given individual, and a number of such variables could easily suffice to identify an individual. ICPSR also gives guidance as to what should be done with indirect identifiers [40]:

> **Treating indirect identifiers.** If, in the judgment of the principal investigator, a variable might act as an indirect identifier (and thus could be used to compromise the confidentiality of a research subject), the investigator should treat that variable in a special manner when preparing a public-use dataset. Commonly used types of treatment are as follows:
>
> - Removal – eliminating the variable from the dataset entirely.
> - Top-coding – restricting the upper range of a variable.
> - Collapsing and/or combining variables – combining values of a single variable or merging data recorded in two or more variables into a new summary variable.

- Sampling – rather than providing all of the original data, releasing a random sample of sufficient size to yield reasonable inferences.
- Swapping – matching unique cases on the indirect identifier, then exchanging the values of key variables between the cases...
- Disturbing – adding random variation or stochastic error to the variable.

These all seem like reasonable things to do, but again there are no objective rules on when each of these things should be done or to what extent. Collapsing variables, sometimes called "binning," is the practice of combining many possible values of a variable into a single category (e.g., replacing an exact address with simply the state of residence). The document goes on, however, to give the example of changing a state of residence in a database to simply a region like "south." While a region does seem less useful to an attacker than a state, a state is already a reasonably generic variable, not at all personally identifiable, and a region certainly still adds *some* ability to narrow down who might be associated with each row.

In general, there is no guidance given for how much needs to be done to a given data set before it is considered anonymized. It is suggested that researchers looking to publish data consult with ICPSR staff or other experts who presumably have more experience and better judgment in making these decisions. But at the end of the day, a judgment call must be made, and those decisions are made in an informal way based on past experience and a reasonable guess at what sort of data a potential adversary might have to work with and what sort of time or expertise they are willing to commit to identifying individuals.

This is not to say that this type of anonymization is useless – clearly it makes it substantially harder to identify individuals in the database and at least dissuades

the casually curious from bothering to do so, but this is not all that is promised when this anonymization is used. As mentioned earlier, the ICPSR guidelines suggest strong promises be made to participants that potentially identifying information will be removed. In fact, this sort of anonymization has also been adopted as the legal standard in federal regulation. For example, federal regulations on the use of medical information states that privacy protections do not apply to the release of "de-identified" data [72]. To qualify as de-identified, it is sufficient that 18 specified elements (mostly things that would be classified above as direct identifiers) have been removed and that the data owner "not have any actual knowledge that the information could be used ... to identify an individual." It is clear that anonymization is being used with the expectation that it protects against a dedicated attacker, but no such claim is justified.

2.3 Attacks

It seems clear to us that these anonymization methods are not sufficient to guarantee privacy. Knowing that a friend participated in a study, a curious individual could look at the released database and use information that is included and that the curious individual knows (say, occupation, number of children, approximate age, etc.) find a row that must be associated with their friend. Unfortunately, this is not mere speculation. Supposedly anonymized databases have been attacked many times, both by academics looking to make a point and by journalists and other members of the public. Here we review some of the most famous attacks.

AOL: In July of 2006, AOL released a database of 20 million search queries from more than 650,000 members in an attempt to aid in academic research [55]. The records were stripped of names, IP addresses, and other account information, but the searches themselves were released. Those searches frequently included locations around where users lived, their family members, and other information that could easily be used to identify them. The New York Times quickly identified the searches of Thelma Arnold of Lilburn, Georgia [2]. Various blogs claimed to have identified other users. Websites were set up mocking the search terms of some users, and at least one play was written based on the implied life of an individual [63]. Following the release, the CTO of AOL resigned and two other employees were fired [43]. A class action lawsuit was also filed against AOL on behalf of those whose privacy was violated.

Massachusetts GIC: In this case, medical records from a database held by the Massachusetts Group Insurance Commission (GIC) were released for academic research purposes. The released data retained the ZIP code, birth date, and gender of each patient. (This was thought to be acceptable for anonymization.) Latanya Sweeney cross-referenced these records with voting registration files (which are publicly available) for Cambridge, Massachusetts and was able to identify the medical records of a large number of people, including the current governor of Massachusetts, who had insisted that the data release was not a privacy concern.

Netflix Prize: In 2006 Netflix began what it called the Netflix Prize. The idea was to improve the recommendation system on the Netflix website. Netflix offered $1 million to the first team that, based on past movie ratings, could predict the rating

a customer would give a future movie with at least 10% greater accuracy than Netflix's existing algorithm. The prize was successful, eventually being awarded to the team "BellKor's Pragmatic Chaos." Netflix was able to get a an improved prediction algorithm at a very small cost. At the same time, substantial general-interest academic research was done. It seemed in general like an excellent outcome for all involved.

However, there were privacy concerns raised by the prize. To facilitate the competition, Netflix released a database of 100 million movie ratings assigned by 480,000 subscribers. The database did not include names and other obviously-identifying information. Narayanan and Shmatikov studied connections between the Netflix data and data publicly available in the Internet Movie Database (IMDb) [57]. They found that with high certainty, they could identify users with a very small amount of outside information. (99% could be identified with only 8 known movie ratings, 2 of which could be wrong, and with dates of movie ratings having 2 weeks of possible error.) This is a concern, because while IMDb ratings are public, they could now be connected to Netflix ratings, which are private (and which might imply, for example, information about a user's politics, religion, or sexual orientation). In fact, movie rental information is well-protected by US law due to an incident where Supreme Court nominee Robert Bork's rental history was obtained by a reporter. A class action lawsuit was filed and settled out of court [69]. A planned second prize for further improvements was canceled due to privacy concerns [38].

GWA Studies: In the field of genetics research, genome-wide association (GWA) studies are a common tool meant to help identify single nucleotide polymorphisms

(SNPs) that are associated with particular diseases. Genotyping thousands of individuals as needed for these studies is expensive, but data can be transfered and combined with data from other studies, so efforts were made to make data available publicly. In order to mask individuals' genetic information often only tables of allele frequencies were shared. However, Homer et al. [37] show statistical methods for using only this summary data to tell whether or not a particular individual was included in the study (which can, by implication, reveal whether or not they have a particular disease). As a result of this attack NIH removed aggregate statistics and results of GWA studies from open-access databases, instead requiring individual approval of each researcher before access was granted. This substantially slows research and makes it cumbersome or impossible for researchers, particularly those new to the field, to get access to needed data.

2.3.1 Lessons Learned

Having seen these attacks, several things should be clear. The first is that the lack of obvious ways to link data to individuals is not a sufficient condition to guarantee privacy. While in retrospect the AOL data release, and arguably the GIC data release as well, was clearly flawed, at the time it was seen as lacking foreseeable attacks. More recent guidelines like those of ICPSR would now find those data releases unacceptable. (Attacks like this are probably why ZIP code is listed as a direct identifier that needs to be removed, even though it is not individually specific in the way other direct identifiers are.) But by relying on judgment and experience to

determine what identifiers are unacceptable and how much manipulation is needed to render them safe, these guidelines will always have trouble anticipating new, innovative attacks. Certainly the Netflix Prize and GWA study data releases would have been declared acceptable by ICPSR's standards. The cryptography community has learned from decades of experience that ruling out known attacks is not sufficient to guarantee safety. Instead, guarantees must be made so that they rule out all possible attacks, even those not anticipated at the time.

In order to do this, it is necessary to use more formal requirements. What must be done to data in order to protect privacy must be rigorously defined, and that requirement must then be defended as accomplishing the desired goal. The bulk of the present work is dedicated to this endeavor.

2.4 k-Anonymity and its Enhancements

Seeing the attacks discussed above, researchers responded by seeking formal requirements that could guarantee privacy was protected. Sweeney, who had been responsible for the GIC attack, and Samarati proposed k-anonymity [67, 70]. It is easiest to understand k-anonymity when remembering that it was proposed as a response to the type of attack used against the GIC data, specifically the use of public databases with some of the same information to link data to specific individuals.

We let x represent a particular database, with rows x_i representing the set of information associated with the i^{th} individual. Furthermore, k-anonymity uses the concept of a *quasi-identifier*. Quasi-identifiers include straightforward identifiers

like name and social security number, but also things like birth date, gender, ZIP code, and other information that could in combination be used to isolate the record of a given individual. Let *quasi* be a function that takes as input a row x_i of the database and outputs the part of the row (say, a subset of attributes) that consists of quasi-identifiers. We then formally define k-anonymity.

Definition 2.1 *A released database x is* k-anonymous *if for each row x_i there exists a subset of database rows $s \subset x$ with $|s| \geq k$ such that for all $x_j \in s$ we have $quasi(x_i) = quasi(x_j)$.*

It should be clear right away that k-anonymity does indeed rule out reidentification attacks based on the quasi-identifiers. If k rows all have exactly the same quasi-identifiers, then quasi-identifiers cannot be used to distinguish which of those rows is associated with a given individual. The definition also provides a quantitative standard for how much the specificity of various fields needs to be degraded before information can be released, parameterized by a value k that in a reasonably understandable way quantifies the amount of privacy being provided.

There are, however, several important criticisms of k-anonymity. The first is that it assumes the data owner knows which fields are quasi-identifiers that the adversary could use. This assumption was clearly stated in the original defining work [70], but it is a fundamental weakness of the definition. It assumes the owner knows what outside information, which we call *auxiliary information*, is available to an attacker, but such information is very hard to predict. For example, k-anonymization on the quasi-identifiers of ZIP code, birth date, and gender would

have prevented the linkage attack carried out by Sweeney. In particular, the linkage could only at best link each individual to a set of k rows, one of which must be the true row with that individual's data. However, Governor Weld, whose data was identified, probably had released some limited information about his health as part of his campaign. This data could have been used to figure out which of the k rows was truly associated with the governor. Similarly, friends and family of private citizens often know some basic information about each other's health. A life insurance company considering taking on a new customer usually has the results of a physical examination and could use these to identify a record in the database, giving them access to more (private) medical information.

The failure to anticipate available auxiliary information was also part of what led to the Netflix attack. Netflix probably did not anticipate attackers having access to a database like that of IMDb where many Netflix users had listed many of their movie preferences. In fact, such databases of auxiliary information might be made available *after* the release of the initial privatized database. The lesson here is that *every* variable is a quasi-identifier.

Of course, one could simply use k-anonymity with all variables considered quasi-identifiers. It would limit the usefulness of the data (though some usefulness would usually remain), but it would certainly be a plausible definition. This does not, however, deal with the second criticism, which is that reidentification is not the only concern. It is possible, for example, that the row representing the governor was limited only to a set of size k, but that all of those rows represented individuals with a particular serious medical condition. This is called a *homogeneity attack* and

constitutes leakage of sensitive information, even though the attacker who learned the information still does not know which row represented the governor.

In order to prevent homogeneity attacks, Machanavajjhala et al. [51] propose ℓ-diversity, a restriction of k-anonymity that requires that the set S_i of rows with identical quasi-identifiers contain a diverse (parameterized by ℓ) variety of values in sensitive atributes. (Diversity can be measured in a variety of ways.) This reduces the problem, but does not eliminate it. With reasonable levels of diversity, the distribution of sensitive values will still differ from the distribution that occurs in the larger population, and as a result an adversary can learn that certain sensitive values are more or less likely.

Perhaps more problematically, ℓ-diversity can only be applied when sensitive values and quasi-identifiers are distinct sets. If, as discussed previously, the attacker knows some (unknown) partial information about the governor's medical history, the definition cannot be satisfied, since it would require both that the medical information within each block be identical between rows, and also that it be diverse.

This line of work continues, with k-anonymity enhanced in various ways (e.g., [48]). The same fundamental weaknesses run through the entire line of work, however. It is continually assumed that only certain variables are useful for identification. More importantly, there is never a definition with an associated theorem that proves, under general assumptions, that the information an adversary could learn about an individual is limited. Instead, analysis generally assumes that the attacker is using a linkage attack similar to that used against the GIC data. The GWA study attack, for example, shows that privacy can be violated even when nothing

remotely resembling reidentification has occurred. (For a more detailed critique of these definitions, see Domingo-Ferrer and Torra [20].)

2.5 The Query Setting

Our discussion so far has dealt with the data release setting. That is, the output being released takes the form of a database. The rows might be modified, with some values deleted or altered, but one can still talk about rows in the output and those rows still have some connection to the individuals whose data was used to generate them. We now move to a more general setting.

While we generally talk about the database, denoted x, as having rows, all that really matters is that which information is associated with which individual is well-defined. (We use x_i to mean the data associated with individual i, and we will sometimes refer to x_i as the "i^{th} row." n is the size of the database, and $\mathcal{U}$ is the universe of possible values for a row, so $x_i \in \mathcal{U}$ and $x \in \mathcal{U}^n$.) Instead of releasing a modified database, we release the answer to some query. This represents an interactive relationship between the user and the database owner. The user submits a query function F, frequently referred to as a *mechanism*, and the database owner runs the function locally, sending $F(x)$ back to the user.

This setting is more general, as it allows for the discussion of summary statistics and other more concise information. Such information is more limited and therefore much easier to release in a private way. We stress that this setting is not actually narrower — the query in question *could* ask for an anonymized version of

x. There has in fact been work on creating synthetic data that shares the important properties of the original database [9]. Nevertheless, in practice releasing such output turns out to be prohibitively hard in most settings and research has focused on queries that are meant to do meaningful analysis and output useful summary statistics, rather than queries meant to output anonymized data that can be put to a variety of uses.

2.6 Mechanisms in the Query Setting

Many mechanisms have been proposed that attempt to allow various queries to be answered while protecting privacy. Below we discuss some of the most important early[1] methods in this area. For more detail, we refer the reader to the survey by Adam and Wortman [1].

Grouping. Some potential attacks on privacy are quite simple. In particular, queries often ask for some statistic (say, a count or a mean) on a subset of the database that satisfies a certain criterion. If this satisfying subset has only a single row, private information is released. Similarly, if two queries apply to subsets that differ by only one row, that row can generally be inferred by comparing the two outputs. Using more queries allows similar attacks to be carried out while obscuring the fact that such an attack was occurring.

Chin and Özsoyoğlu [14,61] propose the *conceptual model* as a way of creating a private database system. Within this framework, privacy is said to be guaranteed

[1]By "early" we mean prior to differential privacy, which is discussed in Chapter 3.

by the division of the database into (disjoint) *atomic populations*. All atomic populations are either empty or of size at least two, and the subset of the database to which a query applies is required to contain all or none of each atomic population. This means that no two query answers, either directly asked or inferred through a combination of other responses, can apply to subsets of the database that differ in only one row. Unfortunately, this does not truly guarantee privacy in all circumstances. Consider, for example, a sum query that says an atomic population of size two has a sum on one attribute of twice the maximum value allowed for that attribute. This clearly implies that both rows have the maximum value of that variable, violating privacy.

Query restriction. This is a general approach that refuses to answer queries of particular types. In general, it is hard to distinguish this approach from a definition, since it establishes a criterion for what an acceptable query is. A series of query restrictions have been studied, focusing on preventing the type of attack described above. This is done by refusing to answer queries that release information about small subsets of the database [27, 28, 36, 68] or about subsets that have high overlap [19]. Chin and Özsoyoğlu [15] give a method of query restriction that can be used to deal with sum queries.

Data perturbation. Several methods were proposed that added random noise of one form or another to each quantitative value in a database. Traub et al. [71] add a straightforward random variable to the value, while Liew et al. [49] achieve a similar goal by replacing true private values with new values generated from a probability

distribution meant to represent that of the original database. While this does add some uncertainty and prevent *exact* disclosure of the value, the analysis provided is incomplete. For example, in summarizing this work Adam and Wortman [1] say that "perturbing a salary of $150,000 by 3000 would be considered a compromise while at the same time perturbing a salary of $15,000 by 3000 would preserve the confidentiality of the data." It is not clear to us what the basis is for such a statement.

Randomized response. Warner [74] develops *randomized response* as a technique for getting survey respondents to answer questions more honestly. The idea is to give instructions to respondents to let them randomize their answers before providing them to the researcher. For example, when asking for a yes-or-no answer, the interviewer could provide the respondent with a spinner that lands on "true" with some probability and "false" otherwise. The respondent could be asked to privately spin the spinner, and then give the true or false answer to the question according to what the spinner does. Crucially, the overall fraction of the population for which each answer is correct can be estimated accurately based on these responses. Even correlations and other useful properties are maintained to at least some extent. While this method was originally proposed in order to get more useful answers to questions respondents often find embarrassing, it can also be used to protect privacy. In fact, this method is often completely defensible even under differential privacy [44], though when originally proposed for use in this way the arguments made for it were much less rigorous. Importantly, this sort of randomization can be added

retroactively by the database owner, rather than in the original data collection.

Random sampling. Denning [18] proposes a system where queries are answered not on the entire database but on a random subset of the database. Like randomized response, variations on this idea are indeed rigorously defensible [29] but arguments of this sort were not given at the time.

2.7 Definitions in the Query Setting

The attempts at private output discussed in the previous section tend to share some common weaknesses. First, they are all presented without a formal definition of privacy. Instead of proving compliance with a general definition meant to capture privacy as a whole, the arguments for these mechanisms are more ad hoc in nature. Many focus on particular types of attacks, preventing for example queries that only disclose information about a single row. While preventing such an attack is necessary to guarantee privacy, it is not sufficient. These arguments also tend to focus on a very particular idea of what constitutes a privacy violation. It is assumed that the goal of the adversary is the exact determination of an attribute's value, or at least a close approximation. Other disclosures, such as the relationship between two attributes, or between the attributes of two different individuals, are not considered. The lessons taken from the attacks discussed earlier apply here as well. In order to be sure privacy is maintained, formal definitions must be given that protect against all attacks, not just those of a particular type. Analysis of these mechanisms also often assumes that the database owner has complete knowledge of the adversary's

auxiliary information, which is clearly an unreasonable assumption in most cases.

2.7.1 Zero-Information

While the mechanisms discussed above were not shown to be private under any formal definition, there were indeed some definitions proposed in the query setting. Of particular importance is that of Dalenius [16]. While we argue that k-anonymity and related definitions are too weak, we believe Dalenius' definition to be too strong. Dalenius proposed that any release of information was a privacy violation if it was "possible to determine the value [of sensitive information about an individual] more accurately than is possible without access to [the released information]." This is an extremely strong requirement, and certainly sufficient for privacy, but it has since received criticism [26] for considering as privacy violations the release of information that most would not intuitively find to be privacy-violating. The difficultly arises from the fact that some auxiliary information about an individual might already be known to an adversary. For example, consider the researcher investigating a potential link between smoking and cancer. If Bob is publicly known to smoke, then the release of information linking smoking to cancer would alter an adversary's belief about the likelihood that Bob has cancer. This does indeed reveal information about Bob, but most would not consider this a privacy violation. In fact, Dalenius' definition would prevent all meaningful information release of any kind. For any potential fact about a general population, there is theoretically possible auxiliary information that an adversary could have—something like "If that fact is true, Bob

has cancer"—that would make disclosure of that fact a privacy violation.

Because of this, we see the Dalenius definition as being too strong, and we prefer a weaker definition. It is important to stress that we are not seeking a weaker definition simply because Dalenius' definition rules out something we would like to do—impossibility results sometimes represent true impossibility, rather than a flaw in the definition. Rather, we prefer a weaker definition because we think the impossibility results show that the Dalenius definition fundamentally does not capture the intuitive idea of privacy we are attempting to formalize. The definition implies that the privacy of individuals can be violated even when their data is not present. By its reasoning, the discovery that smoking caused lung cancer violated the privacy of everyone on earth, because anyone whose smoking status was known by their friends now had their risk of lung cancer substantially disclosed. Nevertheless, we maintain that the general public does not consider that sort of knowledge gain to be a privacy violation. In fact, the whole enterprise of private data analysis is built on the goal of releasing general information that increases our understanding of the world as a whole. What is needed is a definition that distinguishes specific information about individuals from general information about the population (even if that information does imply something about many or all individuals).

Chapter 3: Differential Privacy

We now move to a discussion of *differential privacy*, the current state-of-the-art definition in private data analysis. Differential privacy has gained such widespread acceptance because it is a simple definition with useful properties, and (more importantly) because it avoids the two biggest problems discussed in Chapter 2. It provides strong, provable guarantees about the protection of private information and at the same time is not so broad as to rule out the learning of general non-individualized information.

In Sections 3.1 and 3.2, we introduce the formal definition of differential privacy and discuss the reasons it has been so successful. In Sections 3.3 and 3.4 we discuss what can and cannot be done under the constraints of differential privacy. Finally, we note some of the most prominent criticisms the definition has faced.

3.1 Definition and Meaning

Differential privacy, informally, requires that the output of the private mechanism be "roughly the same" even if an individual's data was to completely change. Intuitively, if the output looks the same regardless of what value an individual's data might have had, that output cannot be used to infer anything about the individual.

In order to formalize this, we need to define when two outputs, each potentially randomized, count as "roughly the same." The correct notion turns out to be that of (ϵ, δ)-indistinguishability.

Definition 3.1 *Two random variables A and B taking values in the same set are* (ϵ, δ)-indistinguishable *(denoted $A \approx_{\epsilon,\delta} B$) if, for all sets S, we have*

$$\Pr[A \in S] \leq e^{\epsilon} \Pr[B \in S] + \delta \quad \textit{and} \quad \Pr[B \in S] \leq e^{\epsilon} \Pr[A \in S] + \delta.$$

When $\delta = 0$ we often omit it and write $A \approx_{\epsilon} B$.

The differential privacy definition compares the real database x to a database x' where individual i had entered arbitrary other data. We call such a pair of databases "neighboring."

Definition 3.2 *Databases x and x' are* neighboring if they differ only in one row.

We are now ready to define differential privacy.

Definition 3.3 *A (randomized) mechanism F is* (ϵ, δ)-differentially private if for all x and x' differing in only one row, we have

$$F(x) \approx_{\epsilon,\delta} F(x').$$

When $\delta = 0$ we refer to this as simply ϵ-differential privacy.[1]

[1]Differential privacy can be also be defined with x' representing a database with a row removed, rather than changed. These definitions are, barring some minor technical details about whether the size of the database itself is hidden, equivalent.

Differential privacy was first defined under the name *ϵ-indistinguishability* by Dwork, McSherry, Nissim, and Smith [25]. That definition lacked the δ term included above, which was added soon after [23].

Differential privacy is what we will refer to as an "output-based" definition, meaning that it gives a condition on what the output of the mechanism should look like. Output-based definitions tend to be easier to work with, but their interpretation is not always so clear. The alternative is what we will call an "inference-based" definition, meaning a definition that speaks directly to what an adversary could infer about an individual based on the output of the mechanism.

Crucially, differential privacy provably implies a meaningful inference-based definition. The definition is weaker than that of Dalenius: instead of requiring that the adversary can learn nothing as a result of the query output, it requires that the adversary learn *nothing more than would be learned if the individual in question was not included in the database.* That is, information that is inferred about an individual because of facts learned about the general population are not protected. Only information that is specific to the individual cannot be learned.

This distinction aligns reasonably well with the interpretation of "privacy" that most people have. In our hypothetical scenario in which Bob is a smoker and a medical study is released that implies he is at higher risk for cancer, most would not think Bob's privacy was violated. If instead the query in question disclosed that Bob's insurance company is spending a large amount of money on his medical care, an adversary might make a similar inference about Bob's chance of having cancer. However, despite the inferences an adversary could make being very similar, the

second scenario would be seen as a greater privacy violation by most people. We argue that the reason for this is that the inference in the second scenario is specific to Bob and depends on data that he contributed, whereas in the first situation it is a consequence of a general fact about the population and can be determined without access to any information of Bob's.

To formally define the inference-based analogue of differential privacy, (ϵ, δ)-*semantic privacy*, we first use B to represent a distribution over the space of possible databases $\mathcal{U}^n$ corresponding to the adversary's prior belief about the database. To the adversary, the real database x is drawn randomly from the distribution B, and then the output of some mechanism $F(x)$ is released. We use $B|_{F(x)=t}$ to denote the distribution B conditioned on a particular output of the mechanism, and we use x_{-i} to represent the database x with the i^{th} row removed and replaced with a fixed default value. Semantic privacy requires that the belief distribution given the mechanism's output is similar whether or not a given individual's data was included in the database, with similarity measured by standard statistical distance.

Definition 3.4 *F is an (ϵ, δ)-semantically private mechanism if for all prior distributions B and all values of i, with probability $1 - \delta$ over pairs (x, t) where x is drawn from B and t is drawn from $F(x)$*

$$\boldsymbol{SD}(B|_{F(x)=t}, B|_{F(x_{-i})=t}) \leq \epsilon,$$

*where **SD** is statistical distance and is defined for two random variables A and A'*

as

$$\boldsymbol{SD}(A, A') = \max_S(|\Pr[A \in S] - \Pr[A' \in S]|). \tag{3.1}$$

The strength of differential privacy comes from the following theorem, proven by Kasiviswanathan and Smith [45], which shows formally that a mechanism satisfying differential privacy also (with a loss in parameters) satisfies semantic privacy.[2] This means that an adversary cannot, regardless of his prior beliefs, learn any more from the mechanism's output than he could have learned without the data of the individual in question being included.

Theorem 3.1 *If a mechanism is* (ϵ, δ)*-differentially private then it is also* (ϵ', δ')*-semantically private, where* $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{\delta}$ *and* $\delta' = O(n\sqrt{\delta})$.

This theorem shows why differential privacy rests on fundamentally firmer ground than k-anonymity and other definitions. Instead of focusing on reidentification or some other particular step on the way to learning about an individual, differential privacy makes direct guarantees about what an adversary can learn about the individuals represented in the database. We emphasize that privacy is guaranteed against all choices of distribution B. In particular, this means that whatever auxiliary information the adversary possesses, they will still learn no more than they would have learned if the a given individual's data was not included.

[2]The reverse is also shown, that any semantically private algorithm is differentially private, but that is not crucial to our current discussion.

3.2 Properties

There are several important properties proven to apply to differential privacy. Some of these are of interest mainly because they support the definition's claim to capture the true meaning of "privacy." Others are important because they are useful technical tools when proving certain mechanisms are private, or because they allow such tools to be used with less restriction in the real world. The most important of these properties are discussed below.

Resistance to post-processing. If a mechanism $F(\cdot)$ is (ϵ, δ)-differentially private, then for any (randomized) function G, $G(F(\cdot))$ is also (ϵ, δ)-differentially private. This means that no amount of computational work can make secrets appear from an output that previously met the privacy criterion. The lack of such a property would be an extremely strong argument against a privacy definition.

Composition over multiple queries. If a series of (adaptively chosen) differentially private queries is considered to constitute one larger query, that larger query is also differentially private. It has ϵ and δ values equal to the sum of the corresponding values of the individual queries. This means first of all that large, complex investigations can be easily analyzed, as long as their smaller, simpler components are private. It also means that database owners need not keep track of any information other than the total ϵ and δ values of all the queries they have allowed to be run on a database. In particular, a researcher can be allotted a "privacy budget" based on how much of the database's ability to be queried they are allowed to use

up, and that researcher can freely make any set of queries that fit within the budget.

Composition over groups of individuals. Differential privacy elegantly distinguishes between information about a population as a whole and information unique to an individual, as discussed previously. Of course, there is information that is in between the two extremes of that continuum. A particular disease might run in a family, for example. Information that is tied to small numbers of individuals intuitively strikes most as more private, similar to information tied to a single person. Information tied to large groups of individuals (a disease being very common in a particular city, for example) doesn't seem as private. However, there is no clear line. Differential privacy deals well with this issue. One can create a "group privacy" definition in which the differential privacy criterion is required to hold when the data of n people (instead of just one) is changed. If a query is (ϵ, δ)-differentially private for individuals, then it has $(n\epsilon, n\delta)$-group privacy for groups of size n. This means that how much information is protected gradually degrades as the size of the group that information is tied to increases.

3.3 Differentially Private Mechanisms

A wide variety of differentially private mechanisms have been invented in order to approximate any number of non-private queries. However, many of these mechanisms have at their root the application of a couple simple ideas/mechanisms that apply very broadly. Because of their importance, we present these mechanisms in full. Following this, we discuss a variety of other mechanisms that have been found,

including private versions of machine learning algorithms.

3.3.1 Sensitivity-Based Noise

If the non-private query q being approximated has output in $\mathbb{R}^d$, we can create a private version by adding random noise. We measure closeness in $\mathbb{R}^d$ using an L_p norm.

Definition 3.5 (L_p-norm) *The L_p-norm of a vector $\mathbf{v} \in \mathbb{R}^n$, denoted $||\mathbf{v}||_p$, is defined as*

$$||\mathbf{v}||_p \stackrel{\text{def}}{=} (|v_1|^p + |v_2|^p + \ldots + |v_n|^p)^{1/p}$$

for $p \in \mathbb{N}^+$, where v_i is the ith coordinate of $\mathbf{v}$. (We do not deal with the L_0 norm in this work.) We also allow $p = \infty$, where

$$||\mathbf{v}||_\infty \stackrel{\text{def}}{=} \max(|v_1|, |v_2|, \ldots, |v_n|).$$

The *sensitivity* of a function is a measure of how much its output can change as the database it takes as input changes.

Definition 3.6 (Sensitivity) *Say that q is a deterministic function on databases with output in $\mathbb{R}^d$. The L_p sensitivity of q, $\mathsf{sen}_p(q)$, is defined as*

$$\mathsf{sen}_p(q) = \max_{x,x'} \|q(x) - q(x')\|_p$$

where x and x' are databases differing in only one row.

Dwork et al. [25] showed that a private mechanism approximating q can be constructed by adding Laplacian random noise that is proportionate to $\mathsf{sen}_1(q)$.

Theorem 3.2 *Take q, a deterministic function on databases with output in $\mathbb{R}^d$. Let $\mathsf{Lap}(b)$ be a random variable from a Laplace distribution centered on 0 with parameter b. Let $F(x) = q(x) + (\mathsf{Lap}(\mathsf{sen}_1(q)/\epsilon))^d$. Then $F(\cdot)$ is ϵ-differentially private.*

While this method works for any query, it is useful primarily for low-sensitivity functions, a reasonably large category. For any predicate, for example, a count of how many rows satisfy that predicate has sensitivity 1. Many other queries, including correlations between boolean variables, can be reduced to a couple such predicate counts. A wide range of mechanisms have been designed with this general method as the primary tool.

It is also possible to achieve similar results using Gaussian rather than Laplacian noise. The following result was proven by Dwork et al. [23].

Theorem 3.3 *Take q, a deterministic function on databases with output in $\mathbb{R}^d$. Let $\mathsf{Norm}(\mu, \sigma)$ be a random variable from a Gaussian distribution centered on μ with standard deviation σ. Let $F(x) = q(x) + (\mathsf{Norm}(0, \mathsf{sen}_2(q)\sqrt{2\ln(2/\delta)}/\epsilon))^d$. Then $F(\cdot)$ is ϵ, δ-differentially private.*

Note that while the constant multiple needed on the Gaussian noise is greater than that needed for Laplacian noise, the sensitivity used is the L_2 sensitivity, which will in general be lower. Gaussian noise is a good example of why allowing $\delta > 0$ is beneficial. No amount of Gaussian noise would achieve privacy with $\delta = 0$. (This

is because the logarithm of the derivative of the density function for the Gaussian distribution becomes large at the tails, rather than remaining constant as it does for the Laplacian distribution.) However, a very, very small δ is sufficient to allow privacy [23].

3.3.2 Exponential Mechanism

The exponential mechanism is a general mechanism that, unlike sensitivity-based noise, can be applied to functions with an output not made up of real numbers or where small amounts of noise might cause substantial harm to utility. Instead, we assume a real-valued utility measure u. $u(x, y)$ is a measure of the utility the user gains when database has value x and the mechanism outputs y. The sensitivity of a given utility measure u is $\max_{x,x',y} \|u(x, y) - u(x', y)\|$. McSherry and Talwar [53] prove the following theorem.

Theorem 3.4 *Let $F(\cdot)$ be a mechanism where the probability of outputting y on input x is proportional to $e^{-\epsilon u(x,y)/2}$. Then F is $\epsilon \cdot \mathsf{sen}(u)$-differentially private.*

This mechanism would be extremely useful if not for the fact that its running time is in general exponential. It is a very interesting feasibility result, but it is only practical for a specific goal if calculating the relevant probability distribution can be done in a much more efficient manner than is known in general.

3.3.3 Other Mechanisms

While not used in this thesis in a way that requires technical detail, several other private mechanisms are worth mentioning.

Smooth sensitivity. Sensitivity is a worst-case measure. A query calculating the median of a list of numbers between 0 and 100, for example, has sensitivity of 100, because there is a possible case where altering one number changes the median from 0 to 100. It is therefore natural to look for a way to add noise proportional to the "local" sensitivity — that is, the maximum change an alteration in a single row could produce *in the current database*. This is, however, not generally acceptable, because the amount of noise added could itself disclose information. (In particular, it could leak the local sensitivity, which is not a private output.) Nissim et al. [58] found a way around this problem through the introduction of *smooth sensitivity*. The idea here is to produce an upper bound on the local sensitivity that has low sensitivity (and can therefore be approximated privately), and use that upper bound to add noise. The result is a substantially more accurate for most of the databases that would occur in realistic settings.

Propose-test-release. Dwork and Lei [24] give another method of dealing with local sensitivity. Here the algorithm uses a proposed bound on the local sensitivity. A (differentially private) test is performed to check whether the particular input database does indeed have local sensitivity below the bound. If it does, an answer is returned. Otherwise, the algorithm simply outputs $\perp$. On arbitrary input databases, an answer of $\perp$ is very common, but given some statistical assumptions

about the distribution of likely databases average utility can be quite good. They use this approach to convert several measurements from the field of *robust statistics* (i.e., statistics designed to be minimally influenced by a few arbitrary outliers) into differentially private algorithms. This includes algorithms for inter-quartile range, median, and linear regression.

Machine learning. There has been substantial interest in developing private versions of learning algorithms. The work in this area falls into two broad categories. The first is a series of general results and broad statements about what can and cannot be learned privately. Kasiviswanathan et al. [44] begin this line of work by showing an algorithm that can learn any PAC-learnable concept class. Unfortunately, this algorithm runs in exponential time and is limited to data drawn from finite, discrete domains. Chaudhuri and Hsu [11] show that when the data is drawn from a continuous domain, there can be very simple hypothesis classes for which private learning is impossible. The generic learner also requires a greater sample size than the non-private learning algorithm. Beimel et al. [4,5] further characterize the sample complexity of private learning and show a distinction between proper and improper learners.

Apart from these general results, a variety of work has focused on particular learning algorithms. For example, private algorithms have been found for learning parity [44], single points [4,5], support vector machines [66], decision trees [8], logistic regression [12,75], and linear regression [24,75]. These algorithms vary in the amount of noise that must be added to ensure privacy.

3.4 Lower Bounds on Differentially Private Mechanisms

It is natural to ask whether lower bounds can be shown to limit the accuracy that can be achieved with differential privacy, and there has been a great deal of work in this direction. The most important result is also the most obvious: the need to randomize responses to a query is unavoidable. Because of the importance of this statement, we formalize it.

Theorem 3.5 *Let F be a deterministic function on a database x of size n. If F is (ϵ, δ)-differentially private for any (finite) values of ϵ and δ, then F is a constant function. That is, there exists c such that $F(x) = c$ for all $x \in \mathcal{U}$.*

More detailed results exist for particular types of queries. For example, *counting queries*, which output the number of rows that satisfy a given condition, have been studied at great length. (For example, see [10, 17, 34].) We forgo a more detailed discussion here because the relevance to this thesis is limited, except to the extent that the existence of such results is part of the motivation for seeking a weakened version of differential privacy.

3.5 Criticisms of Differential Privacy

Several critiques have been made of differential privacy. Here we mention several of the most common and most relevant to our current work.

Parameter choice is arbitrary. Critics sometimes complain that the differential privacy definition says nothing about what is a "good" choice of ϵ (and to a lesser

extent δ). To the supporters of the definition, this is a feature, as the choice is fundamentally a question of policy. Nevertheless, it is sometimes difficult to evaluate research results without knowing what values of ϵ are realistic. Differential privacy has seen limited use in practice, so there are not many instances of policymakers choosing a parameter. One useful data point is an investigation by Chin and Klinefelter [13] into the practices of Facebook. While not publicly confirmed, the number of users reported to potential advertisers as matching a set of selecting criteria seems to include some amount of random noise. Chin and Klinefelter argue that this noise is consistent with the use of differential privacy and estimate that $\epsilon \approx 0.181$.

Differential privacy assumes independence of rows. It is often claimed that the guarantees of differential privacy rest on the assumption that all rows are independent of each other (e.g., [73]). In fact, Kifer and Machanavajjhala [47] even give a "proof" of this claim. Whether the claim is true or not depends on what underlying idea of privacy one is assuming. Differential privacy's supporters make a distinction between individual-level data, which should be protected, and general information about the population, which need not be. This means something like a correlation between smoking and cancer can be freely released. Those saying differential privacy assumes independence of records are implicitly assuming a Dalenius-style definition where any inference about an individual is a privacy violation. As a result, releasing a correlation is a privacy violation. The only way differential privacy protects privacy in this Dalenius-like sense is if no such correlations exist and rows are all independent. We reject the Dalenius understanding of privacy and so reject this

criticism, though we note that it is not *wrong* in an objective sense, but rather true only if one is working from an understanding that already is fundamentally at odds with the motivation for differential privacy.

Differential privacy limits utility. The biggest criticism of differential privacy, largely coming from outside of computer science, is that it does not allow sufficient utility to be gained from the data. Ohm [59] decries the need for an interactive setting, where analysts face delays before queries are answered and cannot look over the data for patterns without specifying a particular query. The delay can be reduced or eliminated with automatic systems like PINQ [54] that answer classes of private queries without the need for human validation. The inability to look at anonymized data, however, is not an a priori limitation being imposed. As stated previously, the query setting allows the data release setting as a special case. The reason data release is not allowed in practice is that releasing data that is reasonably accurate without violating the definition is extremely difficult.

That critique then merges with a more general argument, which is simply that the utility achievable with differential privacy is too low. There are many things that can be done with great accuracy, but of course many queries require prohibitive amounts of noise when run on reasonably-sized databases. In some cases, this is just a matter of more research being needed, but sometimes there are lower bounds or other causes for pessimism. However, this is not on its own a valid criticism of the definition. It is entirely possible that many queries just *cannot* be answered accurately while protecting privacy. That would be unfortunate, but it would not

be cause to reject the definition. In order for this criticism to be complete, it must be argued that the definition is *unnecessarily* strong. That is, one must show that it can be weakened in a way that avoids undermining the convincing argument that the definition truly protects privacy. It is this search for an acceptable weaker definition that motivates the rest of this thesis.

Chapter 4: Computational Differential Privacy

Differential privacy protects against *any* attack. Some of these attacks are extremely simple, consisting largely of a single join over two databases. Others are more complex, using careful statistical analysis and techniques from machine learning. It is entirely possible that some of the attacks being prevented are already infeasible as a practical matter. In particular, attacks against some mechanisms are clearly *computationally* infeasible. Consider, for example, the following examples.

- *F releases the encryption of the database.* This has minimal utility, but clearly seems to protect privacy.

- *F computes a standard differentially private output but uses pseudorandom noise.* Since such noise is computationally impossible to distinguish from truly random noise, this output would be effectively identical to that of the private mechanism for anyone who saw it.

- *F implements a differentially private mechanism using secure multi-party computation (MPC).* Because MPC is proven to disclose nothing but the output of the computation, this seems equivalent to a simple release of the same output.

These mechanisms all seem acceptable. Some also have potential practical benefits. Using MPC, for example, could allow the databases of several hospitals to be queried as if they were a single, large database, without the hospitals having to trust each other and combine the data into a single centralized database. However, differential privacy is a purely information-theoretic notion, and would find all of these mechanisms unacceptable. It therefore seems natural to formalize a *computational* variant of differential privacy.

Mironov et al. [56] formalize just such a definition. In fact they provide several definitions and explore the relationships between them. One would hope that by considering a relaxed definition we can circumvent limitations or impossibility results that arise in the information-theoretic setting in the same way that computational security notions for encryption allow bypassing known bounds for perfectly secure encryption. Initial results [52, 56] showed that this is indeed the case in the *two-party* setting where the database is partitioned between two parties who wish to evaluate some query over their joint data. Specifically, McGregor et al. [52] show a strong separation between the accuracy that can be obtained when using differential privacy as opposed to using *computational* differential privacy.

McGregor et al. [52], however, leave open the analogous question in the more standard *client/server* setting where a server holds the entire database on which a client may pose queries. Indeed, they explicitly remark [52, Section 1]:

> *[Our] strong separation between (information-theoretic) differential privacy and computational differential privacy . . . stands in sharp contrast*

with the client-server setting where... there are not even candidates for a separation.

It is this question we address in this chapter. The results presented here first appeared in TCC 2011 [32].

4.1 Summary of Our Results

There are (at least) two notions of computational privacy that can be considered: IND-CDP and SIM-CDP. These notions are introduced in [56], where it is shown that any SIM-CDP mechanism is also IND-CDP (the other direction is not known); thus, SIM-CDP is a possibly stronger definition. (Mironov et al. also define the notion of $\text{SIM}_{\forall\exists}$-CDP but this notion is equivalent to IND-CDP.) We review these definitions in Section 4.2.

There are two measures one could hope to improve upon when moving from the setting of (statistical) differential privacy to the setting of computational differential privacy: the best possible *utility* (or *accuracy*) that can be achieved, and the *efficiency* of implementing a mechanism that achieves some level of utility. With respect to the definitions given by Mironov et al., it is not hard to see that the best achievable utility cannot be improved as long as the utility is an efficiently computable function of the database and the output of the mechanism. (This is an immediate consequence of the SIM-CDP and $\text{SIM}_{\forall\exists}$-CDP definitions, since otherwise the utility function itself serves as a distinguisher.) The interesting question is therefore to look for improvements in the efficiency, e.g., to show that the best

possible utility *for polynomial-time mechanisms* is better in the computational case, or even to show a polynomial factor improvement in the efficiency in moving from one case to the other. Unfortunately, we show two negative results indicating that such improvements are unlikely in certain natural settings:

1. Our first result concerns *black-box* constructions of computationally secure mechanisms from a wide range of cryptographic primitives including trapdoor permutations, collision-resistant hash functions, and/or random oracles. Roughly, we show that for any black-box construction of a computationally private mechanism there exists a corresponding *statistically* private mechanism that performs just as well in terms of both efficiency and utility (with respect to any utility measure).

2. Our main results rules out improvements by *arbitrary* mechanisms, for a specific (but large) class of queries and utility measures. That is, for queries with output in $\mathbb{R}^d$ (for constant d) and a natural class of utilities, we show that *any* computationally private mechanism can be converted to a statistically private mechanism that is roughly as efficient and achieves almost the same utility.

Each result applies to both the IND-CDP and SIM-CDP definitions.

We believe our results represent an important step in understanding the benefits and limitations of computational notions of privacy. Although we show negative results, they may point toward specific situations where computational differential privacy gives some advantage. We leave it as an open question to find utility measures or query classes with respect to which computational differential privacy *can*

help in the client/server setting, or to extend our impossibility results to show that no such improvements can be hoped for.

Limitations of our results. There are several types of queries to which our results do not apply. The most important are queries with outputs that cannot naturally be thought of as tuples of real numbers. This includes, e.g., queries that return classifiers (as in [44]), graphs, or synthetic databases.

Our results also do not apply, in general, to queries that return output in $\mathbb{R}^d$ for "large" d (i.e., d that grows with the security parameter k). In particular, this means that our results are somewhat limited when it comes to analyzing differential privacy of multiple queries. (Note that d queries with outputs in $\mathbb{R}$ can be viewed as a single query with output in $\mathbb{R}^d$.) Our results do apply to any *constant* number of queries. In addition, using composition properties of differential privacy, our results apply to the case where arbitrarily many queries are answered, and all queries are answered independently (i.e., the server maintains no state). However, in some cases it is known that answering many queries at the same time can be done with better privacy than would be achieved by answering each query independently; in such cases our results do not apply.

Our results also hold only for restricted classes of utility functions. We believe our proof could easily be adjusted for most utilities that measure the expected "closeness" in the reals in some natural way. Less standard ideas of utility, however, might not be covered. For example, a database curator could first output a commitment to the database, then answer queries with differential privacy, then give

a zero-knowledge proof that the queries were answered on the original database to which the commitment applied. There might be settings where this commitment and zero-knowledge proof are beneficial, and this use is not ruled out by our results. (In fact, it is certainly possible.)

4.2 Definitions

Where we have previously talked about a single mechanism F, we now talk about a family of mechanisms $\{F_k\}$, where k is a security parameter. We say a family of mechanisms $\{F_k\}$ is *efficient* if the running time of $F_k(x)$ is at most $\mathsf{poly}(|x|, k)$. A family $\{F_k\}$ is *uniform* if there is a Turing machine F such that $F(k, x) = F_k(x)$. The switch to a parameterized family of mechanisms is necessary to consider computational definitions. It is also reasonable in an information-theoretic setting. In particular, we can require (ϵ, δ)-differential privacy where δ is negligible in k.

Definition 4.1 *Let ϵ be an arbitrary function. A family of randomized functions $\{F_k\}_{k\in\mathbb{N}}$ is $(\epsilon, \mathsf{negl})$-DP if there exists a negligible function δ such that each F_k is $(\epsilon(k), \delta(k))$-DP.*

Mironov et al. [56] propose two definitions of computational differential privacy, SIM-CDP and IND-CDP. Roughly, one can view IND-CDP as an "indistinguishability-based" relaxation whereas SIM-CDP is a "simulation-based" notion. SIM-CDP is at least as strong as IND-CDP [56], but the converse is not known. All the definitions can be presented for either uniform or non-uniform adversaries; for

consistency with [56], we give non-uniform definitions here. While we state our results for the case of non-uniform adversaries, our results all carry over to the uniform setting as well.

IND-CDP provides perhaps the most natural relaxation of differential privacy.

Definition 4.2 (IND-CDP) *Let ϵ be an arbitrary function. A family of functions $\{F_k\}_{k\in\mathbb{N}}$ is ϵ-IND-CDP if for every non-uniform polynomial-time $\mathcal{A}$ and every sequence $\{(x_k, x'_k)\}_{k\in\mathbb{N}}$ of (ordered pairs of) polynomial-size, neighboring databases*[1]*, there is a negligible function* negl *such that*

$$\Pr[\mathcal{A}(F_k(x_k)) = 1] \leq e^{\epsilon(k)} \times \Pr[\mathcal{A}(F_k(x'_k)) = 1] + \mathsf{negl}(k).$$

The notion of SIM-CDP requires that there be a statistically private mechanism that is indistinguishable from the mechanism under consideration.

Definition 4.3 (SIM-CDP) *Let ϵ be an arbitrary function. A family of functions $\{F_k\}_{k\in\mathbb{N}}$ is ϵ-SIM-CDP if there exists a family of functions $\{G_k\}_{k\in\mathbb{N}}$ that is $(\epsilon, \mathsf{negl})$-DP and is computationally indistinguishable from $\{F_k\}$. The latter means there is a negligible function* negl *such that for any non-uniform polynomial-time $\mathcal{A}$ and any database x:*

$$\left|\Pr[\mathcal{A}(F_k(x)) = 1] - \Pr[\mathcal{A}(G_k(x)) = 1]\right| \leq \mathsf{negl}(k).$$

In [56] it is required that $\{G_k\}_{k\in\mathbb{N}}$ be ϵ-DP (rather than $(\epsilon, \mathsf{negl})$-DP). Thus our definition is slightly weaker, which makes our impossibility results stronger.

[1] We abuse notation slightly. Elsewhere subscripts refer to a given row of the database, but in this chapter we do not need such notation and instead use subscripts to refer to databases in a given sequence.

We also recall the notion of $\text{SIM}_{\forall\exists}$-CDP, which weakens SIM-CDP by reversing the order of quantifiers in the definition: here, the statistically private mechanism G is allowed to be different for each pair of databases (x, x'). Crucially for our purposes, this definition is known to be equivalent to IND-CDP [56].

Definition 4.4 ($\text{SIM}_{\forall\exists}$-CDP) *Let ϵ be an arbitrary function. A family of functions $\{F_k\}_{k\in\mathbb{N}}$ is ϵ-$\text{SIM}_{\forall\exists}$-CDP if for all sequences of (unordered pairs of) adjacent databases $\{\{x_k, x'_k\}\}_{k\in\mathbb{N}}$ there is a family of functions $\{G_k\}_{k\in\mathbb{N}}$ such that:*

1. *$\{G_k\}$ is ϵ-DP on $\{\{x_k, x'_k\}\}_{k\in\mathbb{N}}$; i.e., for all subsets $S \subset \mathcal{R}$ we have*

$$\Pr[G_k(x_k) \in S] \leq e^{\epsilon(k)} \times \Pr[G_k(x'_k) \in S].$$

2. *$F_k(x_k)$ and $F_k(x'_k)$ are indistinguishable from $G_k(x_k)$ and $G_k(x'_k)$ respectively. Formally, for any non-uniform, polynomial-time adversary $\mathcal{A}$*

$$\left|\Pr[\mathcal{A}(F_k(x_k)) = 1] - \Pr[\mathcal{A}(G_k(x_k)) = 1]\right| \leq \mathsf{negl}(k),$$

and similarly for x'_k.

Thus far we have only discussed privacy but have not mentioned *utility*. In general, we assume a utility measure U that takes as input a database x and the output of some mechanism $F(x)$ and returns a real number. In Section 4.4 we consider a specific class of utilities.

4.3 Limitations on Black-Box Constructions

Here we show that black-box constructions (of a very general sort) cannot help in the setting of computational differential privacy. (We refer the reader to [65] for further discussion and definitional treatment of black-box constructions.) For concreteness, in the technical discussion we focus on black-box constructions from one-way functions, but at the end of the section we discuss generalizations of the result.

Roughly, a fully black-box construction of an ϵ-IND-CDP mechanism from a one-way function is a family of polynomial-time oracle machines $\{F_k^{(\cdot)}\}_{k\in\mathbb{N}}$ such that for every $\mathcal{A}$ and every $\mathcal{O}$ that is one-way against $\mathcal{A}$ it holds that $\{F_k^{\mathcal{O}}\}_{k\in\mathbb{N}}$ is ϵ-IND-CDP against $\mathcal{A}$. It would make sense also to impose a utility condition on the construction (which could be viewed as a correctness requirement on the constructions), but we do not do so here.

Theorem 4.1 *If there exists a fully black-box construction $\{F_k\}_{k\in\mathbb{N}}$ of an ϵ-IND-CDP mechanism from one-way functions, then there exists an $(\epsilon, \mathsf{negl})$-DP family $\{F'_k\}_{k\in\mathbb{N}}$ that is roughly as efficient and such that, for all databases x and utility measures U,*

$$\left|\mathbf{E}\left[U(x, F_k^{\mathcal{O}}(x))\right] - \mathbf{E}\left[U(x, F'_k(x))\right]\right| \leq \mathsf{negl}(k),$$

where the expectations are both taken over the randomness of the mechanism, and the expectation on the left is additionally taken over random choice of a function $\mathcal{O}$.

Proof: The key idea behind the proof is as follows: a random function is one-way

with overwhelming probability [31, 39]; thus, the mechanism $F_k^{\mathcal{O}}$ with $\mathcal{O}$ chosen at random is also ϵ-IND-CDP. Since the construction is fully black-box (and hence relativizing), one-wayness of $\mathcal{O}$ (and hence indistinguishability of the mechanism) holds even for an unbounded adversary as long as the adversary makes only polynomially many queries to $\mathcal{O}$. We construct F'_k by having it simply run F_k as a subroutine, simulating a random function $\mathcal{O}$ on behalf of F_k. This idea is motivated by analogous techniques used in [31].

Let Func denote the set of length-preserving functions from $\{0,1\}^*$ to $\{0,1\}^*$, and let F'_k be as just described. Then for any adjacent databases x, x' and any (unbounded) $\mathcal{A}$:

$$\Pr[\mathcal{A}(F'_k(x)) = 1] = \Pr_{\mathcal{O}\leftarrow\mathsf{Func}}[\mathcal{A}(F_k^{\mathcal{O}}(x)) = 1]$$

and

$$\Pr[\mathcal{A}(F'_k(x')) = 1] = \Pr_{\mathcal{O}\leftarrow\mathsf{Func}}[\mathcal{A}(F_k^{\mathcal{O}}(x')) = 1].$$

Letting OWF denote the event that $\mathcal{O}$ is one-way, we have

$$\begin{aligned}
\Pr[\mathcal{A}(F'_k(x)) = 1] &\leq \Pr\left[\mathcal{A}(F_k^{\mathcal{O}}(x)) = 1 \mid \mathsf{OWF}\right] + \mathsf{negl}(k) \\
&\leq e^{\epsilon(k)} \times \Pr\left[\mathcal{A}(F_k^{\mathcal{O}}(x')) = 1 \mid \mathsf{OWF}\right] + \mathsf{negl}'(k) \\
&\leq e^{\epsilon(k)} \times \Pr[\mathcal{A}(F'_k(x')) = 1] + \mathsf{negl}''(k).
\end{aligned}$$

The second inequality holds since $\{F_k\}$ is a fully black-box construction of an ϵ-IND-CDP mechanism from one-way functions. (Note that, above, $\mathcal{A}$ is not given

access to $\mathcal{O}$ at all.) But the condition that

$$\Pr[\mathcal{A}(F'_k(x)) = 1] \leq e^{\epsilon(k)} \times \Pr[\mathcal{A}(F'_k(x')) = 1] + \mathsf{negl}''(k)$$

for an unbounded $\mathcal{A}$ is equivalent to $(\epsilon, \mathsf{negl})$-differential privacy.

The claim regarding the utility of $\{F'_k\}$ follows by a similar argument. (Note that we do not require that U be efficiently computable.) □

Note that the above proof holds not just for constructions based on one-way functions, but for any black-box construction from a primitive P that can be instantiated with a random object. This includes, e.g., ideal ciphers, collision-resistant hash functions, and trapdoor permutations [31].

4.4 Limitations for Arbitrary Mechanisms

In the previous section we ruled out black-box constructions from general assumptions, but with regard to arbitrary measures of utility and arbitrary mechanisms. Here, we focus on *arbitrary* mechanisms with output in $\mathbb{R}^n$ (for constant n), and a large, but specific, class of efficiently computable utilities. Specifically, we look at utilities based on the L_p norm (see Definition 3.5), but broadened to include things like mean-squared error that are commonly used in statistics. We assume an ideal (presumably non-private) query q represents the ideal answer, and measure utility as closeness to the output q would give.

Definition 4.5 (Average (p, v)-error) *Let $F_k : \mathcal{U}^* \to \mathbb{R}^d$ be a mechanism for*

answering a query $q : \mathcal{U}^* \to \mathbb{R}^d$*. The* average (p, v)-error *(also called the* v^{th} moment of the L_p error*) of this mechanism* ($p > 0, v \geq 1$) *on database* x *is*

$$\sigma_{p,v}(q, x, F_k) \stackrel{\text{def}}{=} \mathbf{E}\left[||F_k(x) - q(x)||_p^v\right].$$

We often refer to the above as "error" rather than "utility"; lower error values are good, whereas lower utility values are bad. We remark that we can handle utility measures beyond the above, as long as they satisfy a technical requirement that follows from our proof. Since we do not currently have any clean way to state this requirement, we do not discuss it further

Given a mechanism $\{F_k : \mathcal{U}^* \to \mathbb{R}^d\}_{k\in\mathbb{N}}$ for answering a query $q : \mathcal{U}^* \to \mathbb{R}^n$, we say *the average* (p, v)*-error of* $\{F_k\}$ *is polynomially bounded* if there is a polynomial err such that, for all x and k, we have

$$\sigma_{p,v}(q, x, F_k) \leq \mathsf{err}(k).$$

Theorem 4.2, below, shows that nothing can be gained by using computational differential privacy rather than statistical differential privacy, as long as we consider mechanisms whose error is polynomially bounded. Before giving the formal theorem statement and proof in the following section, we give an intuitive explanation here.

Let F_k be a polynomial-time ϵ-SIM-CDP mechanism for answering some query $q : \mathcal{U}^* \to \mathbb{R}^d$, where we assume that F_k also has output in $\mathbb{R}^d$ (and d is independent of k). Let $p > 0, v \geq 1$ be arbitrary, and assume the average (p, v)-error of

F_k is polynomially bounded with error bound err. We claim there is an $(\epsilon, \mathsf{negl})$-DP mechanism $\hat{F}_k$ with essentially the same running time[2] as F_k, and such that $\sigma_{p,v}(q, x, \hat{F}_k) < \mathsf{err}(k) + \mathsf{negl}(k)$.

Let $\{G_k\}$ be a mechanism that is $(\epsilon, \mathsf{negl})$-differentially private and indistinguishable from $\{F_k\}$. Such a mechanism is guaranteed to exist by the definition of SIM-CDP. Note that $\{G_k\}$ may be much less efficient than $\{F_k\}$, and may not even be implementable in polynomial time. On the other hand, G_k and F_k must induce distributions over $\mathbb{R}^d$ that are, in some sense, very close. Intuitively, in any "box" in $\mathbb{R}^d$ of noticeable size, the probabilities with which the outputs of G_k or F_k lie in that cell must be roughly equal; if not, the difference in probabilities could be used to distinguish G_k and F_k (since membership in the box can be efficiently tested).

We derive $\hat{F}_k$ by adding a small amount of uniform noise to the output of F_k. Carefully setting the amount of noise to be sufficiently small, we can bound the error introduced in moving from F_k to $\hat{F}_k$. To analyze privacy of the resulting mechanism, we look at the mechanism $\hat{G}_k$ where a small amount of uniform noise is added to G_k. For any particular value a, the probability with which $\hat{F}_k$ (resp., $\hat{G}_k$) outputs a is proportional to the probability that F_k (resp., G_k) outputs a value within a box centered at a. This box is sufficiently big so that $\hat{G}_k$ and $\hat{F}_k$ have similar probabilities of outputting any particular value.

While $\hat{G}_k$ and $\hat{F}_k$ have similar probabilities of outputting any particular value, the small differences could, in principle, compound and become unacceptably large when summed over all values in some set $S \subset \mathbb{R}^d$. To show that such differences do

[2]Specifically, $\hat{F}_k$ runs F_k and adds a random number to its output.

not grow too large, we use the fact that F_k has polynomially bounded error. This allows us to break our analysis into two parts: one focusing on a region S_c "close" to the correct answer $q(x)$, and the other focusing on $S_f = S \setminus S_c$. We show that

$$\left|\Pr[\hat{F}_k(D) \in S_c] - \Pr[\hat{G}_k(D) \in S_c]\right|$$

is small, using the argument discussed above; we also show that

$$\max\{\Pr[\hat{F}_k(D) \in S_f], \Pr[\hat{G}_k(D) \in S_f]\}$$

is small by the polynomial bound on the error. Combined, this shows that for every S, the difference

$$\left|\Pr[\hat{F}_k(D) \in S] - \Pr[\hat{G}_k(D) \in S]\right|$$

is small, as required. Since G_k, and hence $\hat{G}_k$, is *statistically* differentially private, this means that $\hat{F}_k$ is also.

Formal details are given in the following section.

4.4.1 Statement and Proof of the Main Result

We first present a proof that applies to the (stronger) SIM-CDP definition. We then outline the changes needed to prove the result for the case of IND-CDP.

Theorem 4.2 *Fix $p > 0, v \geq 1$. Let $\{F_k : \mathcal{U}^* \to \mathbb{R}^d\}$ be an efficient ϵ-SIM-CDP mechanism whose average (p, v)-error is polynomially bounded by err. Then there is an efficient $(\epsilon, \mathsf{negl})$-DP mechanism $\{\hat{F}_k\}$ with $\sigma_{p,v}(q, x, \hat{F}_k) < \mathsf{err}(k) + \mathsf{negl}(k)$.*

Moreover, $\hat{F}_k$ has essentially the same running time as F_k; specifically, $\hat{F}_k$ only adds uniform noise to F_k.

Proof: Let $\{G_k\}$ be an $(\epsilon, \mathsf{negl})$-DP family of mechanisms that is indistinguishable from $\{F_k\}$. Let negl_1 be a negligible function such that for any non-uniform polynomial-time $\mathcal{A}$ and any database x,

$$\left|\Pr[\mathcal{A}(F_k(x)) = 1] - \Pr[\mathcal{A}(G_k(x)) = 1]\right| \leq \mathsf{negl}_1(k).$$

(Such a function exists by definition of SIM-CDP.)

Since $\{F_k\}$ is efficient, its output must have some polynomial length. We assume that F_k (and hence G_k) give output in fixed-point notation with k bits of precision. Formally, let $\mathbb{R}_k$ be the set

$$\mathbb{R}_k = \{a \in \mathbb{R} \mid \exists j \in \mathbb{Z} : a = j \cdot 2^{-k}\};$$

then we assume that F_k gives output in $\mathbb{R}_k^d$. (More generally, the proof given here works when the precision is any polynomial in k. Moreover, fixed-point notation is not essential; in particular, the proof can be modified for the case when the output of F_k is given in floating-point notation.[3]) For $a \in \mathbb{R}$ and $k \in \mathbb{N}$, define $\lceil a \rceil_k \stackrel{\text{def}}{=} \lceil a \cdot 2^k \rceil \cdot 2^{-k}$ to be the value a "rounded up" so that it lies in $\mathbb{R}_k$.

[3]The proof can also be modified to handle continuous output, though such output is not natural in a computational setting. The output can be modeled as including whichever (polynomial number of) digits the adversary requests, or even as an oracle that allows arbitrary access to digits of the output.

A set $\mathcal{B} \subset \mathbb{R}^d$ is a *box* if it a Cartesian product of closed intervals in $\mathbb{R}$. Abusing notation, we call a sequence $\{\mathcal{B}_k\}$ of boxes $\mathcal{B}_k \subset \mathbb{R}_k^n$ a box as well. The following is an immediate consequence of the SIM-CDP definition (recall the definition requires indistinguishability against non-uniform adversaries):

Lemma 4.1 *For any box $\{\mathcal{B}_k\}$ and any database x:*

$$|\Pr[F_k(x) \in \mathcal{B}_k] - \Pr[G_k(x) \in \mathcal{B}_k]| \leq \mathsf{negl}_1(k).$$

We next define two mechanisms $\{\hat{G}_k\}$ and $\{\hat{F}_k\}$ that are "noisy" versions of $\{G_k\}$ and $\{F_k\}$, respectively. Because we are dealing with discrete rather than continuous values, the definition is more complicated than simply adding uniform noise in some range.

Set $c(k) = \left\lceil \sqrt[4d]{\mathsf{negl}_1(k)} \right\rceil_k$ For $\mathbf{a} \in \mathbb{R}_k^d$, let $\mathcal{B}_{c,k}(\mathbf{a})$ denote the box with radius $c(k)$ (in the L_∞ norm) centered at $\mathbf{a}$; that is,

$$\mathcal{B}_{c,k}(\mathbf{a}) = \left\{\mathbf{b} \in \mathbb{R}_k^d \,:\, ||\mathbf{b} - \mathbf{a}||_\infty \leq c(k)\right\}.$$

Mechanism $\{\hat{F}_k\}$ is defined as follows: $\hat{F}_k(x)$ computes $F_k(x)$, and then outputs a uniform value in $\mathcal{B}_{c,k}(F(x))$. (This is equivalent to adding uniform, independent, discretized noise from $[-c(k), c(k)]$ to each coordinate of $F(x)$.) Mechanism $\{\hat{G}_k\}$ is defined to be the analogous mechanism that adds noise to G instead of F.

$B_{c,k}(\mathbf{a})$ contains $\left(c(k)\cdot 2^{k+1}+1\right)^d$ points and thus, for any x and $\mathbf{a}\in\mathbb{R}_k^d$:

$$\Pr[\hat{G}_k(x)=\mathbf{a}] \;=\; \left(c(k)\cdot 2^{k+1}+1\right)^{-d}\cdot \Pr[G_k(x)\in B_{c,k}(\mathbf{a})]$$

and

$$\Pr[\hat{F}_k(x)=\mathbf{a}] \;=\; \left(c(k)\cdot 2^{k+1}+1\right)^{-d}\cdot \Pr[F_k(x)\in B_{c,k}(\mathbf{a})].$$

Taking $\mathcal{B}_k=\mathcal{B}_{c,k}(\mathbf{a}_k)$ (for an arbitrary sequence $\{\mathbf{a}_k\}$ with $\mathbf{a}_k\in\mathbb{R}_k^d$) in Lemma 4.1, we obtain:

$$\begin{aligned}
&\left|\Pr[\hat{G}_k(x)=\mathbf{a}_k]-\Pr[\hat{F}_k(x)=\mathbf{a}_k]\right| \\
&\quad= \left(c(k)\cdot 2^{k+1}+1\right)^{-d}\cdot\left|\Pr[G_k(x)\in B_{c,k}(\mathbf{a}_k)]-\Pr[F_k(x)\in B_{c,k}(\mathbf{a}_k)]\right| \\
&\quad\le \left(c(k)\cdot 2^{k+1}+1\right)^{-d}\cdot \mathsf{negl}_1(k). \qquad (4.1)
\end{aligned}$$

The above holds for an arbitrary database x, and so it also holds for any adjacent database x'.

$\hat{G}_k$ applies post-processing to the output of G_k, so $\{\hat{G}_k\}$ is also (ϵ,negl)-DP. Let negl_2 be a negligible function such that for all sets S and adjacent databases x and x' it holds that

$$\Pr[\hat{G}_k(x)\in S]\le e^{\epsilon(k)}\times\Pr[\hat{G}_k(x')\in S]+\mathsf{negl}_2(k). \qquad (4.2)$$

Our goal is to prove that $\hat{F}_k(x)$ is *statistically* close to $\hat{G}_k(x)$, for any x, which will then imply the theorem. We have already shown (cf. Equation (4.1)) that the distributions of $\hat{F}_k(x)$ and $\hat{G}_k(x)$ are *pointwise* negligibly close. We need to show that this is true also for arbitrary subsets. To do this, we first use the polynomial error bound on F_k to argue that F_k (and hence $\hat{F}_k$) must put relatively low weight on outputs that are far from the correct output. Formally:

Lemma 4.2 *There is a polynomial w such that, for any x, we have*

$$\sigma_{p,v}(q, x, \hat{F}_k) \leq \mathsf{err}(k) + c(k) \cdot w(k).$$

The lemma follows from the observation that, for any fixed output $b = F_k(x)$, the output $\hat{b} = \hat{F}_k(x)$ satisfies

$$\left|\left|\hat{b} - q(x)\right|\right|_p \leq ||b - q(x)||_p + d \cdot c(k).$$

The proof of the lemma is tedious, and so we defer it to after the main body of the proof is complete.

Fix an arbitrary x. We now show that with high probability the output of $\hat{F}_k(x)$ is close to the true answer $q(x)$. Set $z(k) = \left\lceil \frac{1}{\sqrt[4d]{\mathsf{negl}_1(k)}} \right\rceil_k$, and define

$$\mathsf{Close}_k \stackrel{\text{def}}{=} \{\mathbf{a} \in \mathbb{R}_k^d : ||\mathbf{a} - q(x)||_p^v \leq z(k)\};$$

i.e., these are the points close to $q(x)$. Let $\mathsf{Far}_k \stackrel{\text{def}}{=} \mathbb{R}_k^d \setminus \mathsf{Close}_k$. Because the average error of $\hat{F}_k$ is at most $\mathsf{err}(k) + w(k) \cdot c(k)$, we have

$$\Pr[\hat{F}_k(x) \in \mathsf{Far}_k] \leq (\mathsf{err}(k) + w(k) \cdot c(k)) / z(k). \tag{4.3}$$

Indistinguishability of $\{F_k\}$ and $\{G_k\}$, and the manner in which $\{\hat{F}_k\}$ and $\{\hat{G}_k\}$ are constructed, implies that $\{\hat{F}_k\}$ and $\{\hat{G}_k\}$ are indistinguishable as well. As in the proof of Lemma 4.1, this means that

$$\left| \Pr[\hat{F}_k(x) \in \mathsf{Far}_k] - \Pr[\hat{G}_k(x) \in \mathsf{Far}_k] \right| \leq \mathsf{negl}_1(k).$$

Combining this with Equation (4.3) yields

$$\Pr[\hat{G}_k(x) \in \mathsf{Far}_k] \leq (\mathsf{err}(k) + w(k) \cdot c(k)) / z(k) + \mathsf{negl}_1(k).$$

We now use the above results to relate the probabilities that $\hat{G}_k(x)$ or $\hat{F}_k(x)$ lie within some arbitrary set. The number of points in Close_k is bounded from above by $(z(k) \cdot 2^{k+1} + 1)^d$, since its size is largest (for fixed $z(k)$) when $p = \infty$ and $v = 1$. For any $S_k \subset \mathbb{R}_k^d$, we can thus lower-bound $\Pr[\hat{G}_k(D) \in S_k]$ via

$$\begin{aligned}
\Pr[\hat{G}_k(x) \in S_k] &= \sum_{\mathbf{a} \in S_k} \Pr[\hat{G}_k(x) = \mathbf{a}] \\
&\geq \sum_{\mathbf{a} \in S_k \cap \mathsf{Close}_k} \Pr[\hat{G}_k(x) = \mathbf{a}] \\
&\geq \sum_{\mathbf{a} \in S_k \cap \mathsf{Close}_k} \left(\Pr[\hat{F}_k(x) = \mathbf{a}] - \left(c(k) \cdot 2^{k+1} + 1\right)^{-d} \cdot \mathsf{negl}_1(k) \right),
\end{aligned}$$

using Equation (4.1), which bounds the difference in probabilities between $\hat{F}_k$ and $\hat{G}_k$ pointwise. Continuing, we have

$$\begin{aligned}
&\Pr[\hat{G}_k(x) \in S_k] \\
&\quad \geq \Pr[\hat{F}_k(x) \in S_k \cap \mathsf{Close}_k] - \left(z(k) \cdot 2^{k+1} + 1\right)^d \cdot \left(c(k) \cdot 2^{k+1} + 1\right)^{-d} \cdot \mathsf{negl}_1(k) \\
&\quad \geq \Pr[\hat{F}_k(x) \in S_k \cap \mathsf{Close}_k] - \left(\frac{z(k)+1}{c(k)}\right)^d \cdot \mathsf{negl}_1(k) \\
&\qquad + \left(\Pr[\hat{F}_k(x) \in S_k \cap \mathsf{Far}_k] - \left(\mathsf{err}(k) + w(k) \cdot c(k)\right)/z(k)\right) \\
&\quad \geq \Pr[\hat{F}_k(x) \in S_k] - \left(\frac{z(k)+1}{c(k)}\right)^d \cdot \mathsf{negl}_1(k) - \left(\mathsf{err}(k) + w(k) \cdot c(k)\right)/z(k). \quad (4.4)
\end{aligned}$$

Similarly, we can upper-bound $\Pr[\hat{G}_k(x) \in S_k]$ via

$$\begin{aligned}
&\Pr[\hat{G}_k(x) \in S_k] \\
&\quad \leq \sum_{\mathbf{a} \in S_k \cap \mathsf{Close}_k} \Pr[\hat{G}_k(x) = \mathbf{a}] + \Pr[\hat{G}_k(x') \in \mathsf{Far}_k] \\
&\quad \leq \sum_{\mathbf{a} \in S_k \cap \mathsf{Close}_k} \left(\Pr[\hat{F}_k(x) = \mathbf{a}] + \left(c(k) \cdot 2^{k+1} + 1\right)^{-d} \cdot \mathsf{negl}_1(k)\right) \\
&\qquad + \Pr[\hat{G}_k(x) \in \mathsf{Far}_k] \\
&\quad \leq \Pr[\hat{F}_k(x) \in S_k] + \left(\frac{z(k)+1}{c(k)}\right)^d \cdot \mathsf{negl}_1(k) \\
&\qquad + \left(\mathsf{err}(k) + w(k) \cdot c(k)\right)/z(k) + \mathsf{negl}_1(k). \quad (4.5)
\end{aligned}$$

Equations (4.4) and (4.5) hold for an arbitrary database x, and thus also hold for any adjacent database x'. Substituting into Equation (4.2) and simplifying, we

obtain

$$
\begin{aligned}
&\Pr[\hat{F}_k(x) \in S_k] \\
&\leq \quad e^{\epsilon(k)} \times \Pr[\hat{F}_k(x') \in S_k] \\
&\qquad + \left(e^{\epsilon(k)} + 1\right) \times \left(\left(\frac{z(k)+1}{c(k)} \right)^d \mathsf{negl}_1(k) + \left(\mathsf{err}(k) + w(k) \cdot c(k)\right)/z(k) \right) \\
&\qquad + e^{\epsilon(k)} \cdot \mathsf{negl}_1(k) + \mathsf{negl}_2(k)\,.
\end{aligned}
$$

We show that the additive terms are all negligible. Note first that

$$
\begin{aligned}
\left(\frac{z(k)+1}{c(k)} \right)^d \cdot \mathsf{negl}_1(k) &\leq \left(\frac{\frac{1}{\sqrt[4d]{\mathsf{negl}_1(k)}} + 2}{\sqrt[4d]{\mathsf{negl}_1(k)}} \right)^d \cdot \mathsf{negl}_1(k) \\
&\leq \left(\frac{3}{\sqrt[2d]{\mathsf{negl}_1(k)}} \right)^d \mathsf{negl}_1(k) \\
&\leq 3^d \cdot \sqrt{\mathsf{negl}_1(k)},
\end{aligned}
$$

which is negligible in k (recall d is constant). To bound $\left(\mathsf{err}(k) + w(k) \cdot c(k)\right)/z(k)$, take k large enough so that $w(k) \cdot c(k) \leq \mathsf{err}(k)$ (this is always possible, since c is negligible while err and w are polynomial). We then have

$$
\frac{\mathsf{err}(k) + w(k) \cdot c(k)}{z(k)} \leq 2 \cdot \mathsf{err}(k) \cdot \sqrt[4d]{\mathsf{negl}_1(k)},
$$

which is negligible. We conclude that $\{\hat{F}_k\}$ is $(\epsilon, \mathsf{negl})$-DP. □

We now return to the deferred proof of Lemma 4.2.

Proof: Let Y_k be the set of possible distances between two points in $\mathbb{R}_k^d$; i.e.,

$$Y_k \stackrel{\text{def}}{=} \{y \in \mathbb{R} \mid y = ||\mathbf{a}_1 - \mathbf{a}_2||_p \text{ for some } \mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}_k^d\}.$$

Let $p_{y,k} \stackrel{\text{def}}{=} \Pr\left[y - 2^{-k} < ||F_k(x) - q(x)||_p \leq y\right]$. Then, by the assumption of our theorem,

$$\sigma_{p,v}(q, x, F_k) \leq \textstyle\sum_{y \in Y_k} p_{y,k} \cdot y^v \leq \mathsf{err}(k).$$

We can upper-bound $\sigma_{p,v}(q, x, \hat{F}_k)$ by assuming that the noise added by $\hat{F}_k$ moves the output further away from the correct answer $q(x)$. In the worst case (when $p = 1$), this increases the distance between the output and $q(x)$ by at most $c'(k) \stackrel{\text{def}}{=} d \cdot c(k)$. Therefore,

$$\sigma_{p,v}(q, x, \hat{F}_k) \;\; \leq \;\; \textstyle\sum_{y \in Y_k} p_{y,k} \cdot (y + c'(k))^v.$$

Using Taylor's theorem, $(y + c'(k))^v \leq y^v + v \cdot (y + c'(k))^{v-1} \cdot c'(k)$. Thus, for k sufficiently large it holds that

$$\begin{aligned}
\sigma_{p,v}(q, x, \hat{F}_k) &\leq \textstyle\sum_{y \in Y_k} p_{y,k} \cdot (y^v + v \cdot (y + c'(k))^{v-1} \cdot c'(k)) \\
&\leq \mathsf{err}(k) + \textstyle\sum_{y \in Y_k} p_{y,k} \cdot \left(v \cdot (y + c'(k))^{v-1} \cdot c'(k)\right) \\
&\leq \mathsf{err}(k) + v \cdot c'(k) \cdot \textstyle\sum_{y \in Y_k} p_{y,k} \cdot (y + d)^{v-1},
\end{aligned}$$

using for the last inequality the fact that $c'(k) \leq d$ for k large enough.

If $y \leq d$ then $(y+d)^{v-1} \leq (2d)^{v-1}$, while if $y \geq d$ then $(y+d)^{v-1} \leq (2y)^{v-1}$. As a result, we can bound the expression above as

$$\begin{aligned}
&\sigma_{p,v}(q, x, \hat{F}_k) \\
&\leq \mathsf{err}(k) + v \cdot c'(k) \cdot \sum_{y \in Y_k} p_{y,k} \cdot 2^{v-1} \cdot (d^{v-1} + y^{v-1}) \\
&\leq \mathsf{err}(k) + v \cdot c'(k) \cdot \left(\sum_{y \in Y_k} p_{y,k} \cdot 2^{v-1} d^{v-1} + \sum_{y \in Y_k} p_{y,k} \cdot 2^{v-1} y^{v-1} \right) \\
&\leq \mathsf{err}(k) + v \cdot c'(k) \cdot \left(2^{v-1} d^{v-1} + 2^{v-1} \sum_{y \in Y_k} p_{y,k} \cdot y^{v-1} \right).
\end{aligned}$$

Since $y > 0$, we have $y^{v-1} \leq y^v + 1$. Then:

$$\begin{aligned}
\sigma_{p,v}(q, x, \hat{F}_k) &\leq \mathsf{err}(k) + v \cdot c'(k) \cdot \left(2^{v-1} d^{v-1} + 2^{v-1} \sum_{y \in Y_k} p_{y,k} \cdot (y^v + 1) \right) \\
&\leq \mathsf{err}(k) + v \cdot c'(k) \cdot \left(2^{v-1} d^{v-1} + 2^{v-1} \cdot (\mathsf{err}(k) + 1) \right) \\
&\leq \mathsf{err}(k) + c(k) \cdot \left(2^{v-1} v \cdot d^v + 2^{v-1} v \cdot n \cdot (\mathsf{err}(k) + 1) \right).
\end{aligned}$$

Since err is polynomial and d, v are constants, this completes the proof. □

The case of IND-CDP. A result analogous to the above holds also for the case of IND-CDP. This follows fairly easily using the equivalent formulation of IND-CDP in terms of $\mathrm{SIM}_{\forall\exists}$-CDP. The difference between SIM-CDP and $\mathrm{SIM}_{\forall\exists}$-CDP is with respect to the order of quantifiers, but this has no real effect on our proof. Note, in particular, that our construction of $\{\hat{F}_k\}$ does not depend, either explicitly or implicitly, on $\{G_k\}$.

Chapter 5: Coupled-Worlds Privacy

We now move to another potential weakening of differential privacy. Suppose Facebook were to release the average income of its users—not a noisy version of the average, but its exact value. Or, suppose an Internet dating service were to release exact aggregate statistics about its users' romantic preferences and sexual habits, as does OkCupid [60]. Such disclosures violate differential privacy (for any ϵ) because they are deterministic, but they do not appear to constitute an actual privacy violation. An adversary cannot use the released information to learn anything sensitive about an individual user (or even a small set of users) without unrealistically precise knowledge about the millions of users of those sites. Differential privacy appears to be overkill in these settings: it provides strong privacy guarantees for an individual user *even if* an adversary knows everything about the dataset besides that user's data, but in the scenarios just considered such omniscience is implausible.

The goal of this chapter is to develop rigorous definitions of privacy for statistical databases that allow us to reason about and exploit existing *adversarial uncertainty* about the underlying data. We are driven by several motivations:

- *Better mechanisms:* As discussed previously, relaxing definitions of privacy potentially allows for mechanisms achieving greater accuracy while still meeting

satisfactory notions of privacy.

- *Analyzing existing mechanisms:* A broader goal is to understand what privacy guarantees are achieved by methods in use today (e.g., disclosure-control methods currently used by statistical agencies, or releases that are mandated by law) that were not designed with specific privacy definitions in mind. In some cases, our definitions provide a starting point for making rigorous statements about such methods.

- *Better understanding of the "semantics" of privacy:* Any definitional effort involves translating from natural-language descriptions of privacy to mathematical formulations of the same. We seek to understand the implications of different definitional approaches for the possible *inferences* about sensitive data that an adversary can make based on statistical releases.

The framework we introduce here, *coupled-worlds privacy*, is flexible and admits several instantiations—including one that is equivalent to differential privacy—and we thus view it as a starting point for future work. We also explore a specific instantiation of the framework that we call *distributional differential privacy*, and illustrate its applicability by studying several appealing "noiseless" mechanisms satisfying the resulting definition.

Some previous works have also looked at modeling and exploiting adversarial uncertainty in private data analysis [6, 21, 33, 47], with the most relevant being the work on noiseless privacy [6] and the Pufferfish framework [47]. (Noiseless privacy can be viewed as one instantiation of the Pufferfish framework.) Both can be viewed

as attempts to formalize Dalenius's definition [16]. As we argued in Section 2.7.1, Dalenius's definition is unreasonably strong [22, 26, 46] since it rules out learning global information about a dataset.

In contrast, we start with the premise that learning global information about some population (e.g., a link between smoking and cancer) is *not* a privacy violation. This is, in part, because learning such global information is the main goal of many statistical studies, and in part because it seems counter-intuitive to speak of a violation of a user's privacy that occurs whether or not that user participates in a study (as in the smoking example). This perspective motivates us to define privacy, as in the case of differential privacy, by comparing the effects of a real-world disclosure to a disclosure computed on a "scrubbed" dataset with, e.g., a user's individual data removed. As we discuss further in Section 5.5, this results in definitions very different from those of noiseless privacy or the Pufferfish framework.

Results in this chapter first appeared in FOCS 2013 [3].

5.1 Our Contributions

We now describe our contributions in more detail.

Definitional framework. We give a framework, *coupled-worlds privacy*, for specifying privacy definitions. As an important example instantiation, we consider *distributional differential privacy*, which generalizes differential privacy. Differential privacy can be thought of as requiring that $F(x)$ reveals nothing more about x_i beyond what would be revealed by $F(x_{-i})$ (where x_{-i} denotes the dataset with x_i

removed). Roughly speaking, distributional differential privacy relaxes differential privacy by treating x as a random variable from some distribution in a pre-specified class of distributions Δ, rather than as a fixed value. This means that x_i can be masked by the randomness of the other rows of the database, rather than just by the randomness introduced by the mechanism. (If Δ is taken to be the class of all distributions, this definition is equivalent to differential privacy.)

A bit more formally, our definition requires indistinguishability of the real world in which $F(x)$ is released, and an ideal world in which a simulator releases some function of the "scrubbed" dataset x_{-i}. In each case, the dataset x is drawn from the same distribution in some class Δ specified as part of the definition. Indistinguishability implies, in particular, that the real-world mechanism "leaks" little more than could be inferred from the "scrubbed" dataset in the ideal world, at least under the assumption that one of the distributions in Δ adequately models the true distribution of x and the attacker's auxiliary knowledge (if any).

We prove various properties of definitions within our framework. Although composition does not automatically hold, we show a condition under which it does. We also show that the class of distributions for which a given mechanism satisfies our framework is *convex*. This is a desirable feature (not shared by some previous definitions) since it implies that if a mechanism is private under distributions (i.e., beliefs) $\mathcal{D}$ and $\mathcal{D}'$, then it is also private under a belief that assigns non-zero probability to each of those distributions. Our framework can be instantiated in several ways to yield different definitions. In particular, as in Pufferfish [47], one can tailor the information considered sensitive by appropriate choice of the "scrubbing"

operation applied to the dataset given to the simulator.

In addition to the Pufferfish framework, we are aware of at least two concurrent efforts to generalize differential privacy that share some broad ideas, one by Bhowmick and Dwork [7] and one by Dwork, Reingold, Rothblum, and Vadhan [64].

Inference-based semantics. As a way of justifying our definitional framework, we formalize an intuitive, inference-based notion of privacy in terms of a Bayesian attacker who updates her belief about the dataset X given the output of some mechanism. We show that if a mechanism is private within our framework then, with high probability, an adversary's posterior beliefs in the real world and the ideal world are close. This is analogous to the result of Theorem 3.1 that differential privacy implies semantic privacy.

The inference-based version of our definition provides a more transparent view of the key difference between our approach and that of previous work taking adversarial uncertainty into account [6, 47]. Previous approaches can be seen as requiring an attacker's posterior belief (in the real world) to be close to its prior belief. Here, in contrast, we compare an attacker's posterior belief in the real world to its posterior belief in a hypothetical ideal world involving a "scrubbed" version of the dataset. This results in a more relaxed definition that is arguably more natural; see further discussion in Section 5.5.

Analyses of specific mechanisms. On an intuitive level, there are two different ways to exploit the fact that our definition considers datasets drawn from some distribution rather than a "worst-case" dataset as in differential privacy. The first

is to leverage the uncertainty of the database to avoid adding noise to the output. The second is to argue that the database sampled will, with high probability, satisfy some condition under which privacy holds. We use these ideas to design several natural, "noiseless" mechanisms. These mechanisms will be presented in detail in Chapter 6.

5.2 Background

We begin by presenting an equivalent formulation of differential privacy proposed by Gehrke et al. [30] that utilizes the simulation-based method of defining security that is popular in cryptography. This definition requires that the true output can be (approximately) simulated without access to a given individual's data.

Definition 5.1 *A mechanism F is* (ϵ, δ)-differentially private *if there exists a simulator* Sim *such that for all x and i*

$$F(x) \approx_{\epsilon,\delta} \mathsf{Sim}(x_{-i}).$$

This is easily seen to be equivalent to Definition 3.3: If $F(x) \approx_{\epsilon,\delta} F(x')$ for all neighboring x, x', then a valid simulator is given by the algorithm that inserts an arbitrary entry into x_{-i} and then applies F to the result. Conversely, if a suitable Sim exists then for any two datasets x, x' that differ in the i-th element we have $x_{-i} = x'_{-i}$ and hence $F(x) \approx_{\epsilon,\delta} \mathsf{Sim}(x_{-i}) = \mathsf{Sim}(x'_{-i}) \approx_{\epsilon,\delta} F(x')$.

We expect the fact that the same output can be simulated without access to x_i to mean that $F(x)$ reveals no information that could not be learned from x_{-i}. We can verify this by formulating an inference-based version in the simulation paradigm. Specifically, we require the existence of a simulator Sim such that for all distributions on the database (now a random variable X) and indices i, and with probability $1 - \delta$ over the choice of $t = F(X)$, we have[1]

$$X_i\big|_{F(X)=t} \approx_{\epsilon,\delta} X_i\big|_{\mathsf{Sim}(X_{-i})=t}. \tag{5.1}$$

This formalizes the common interpretation of differential privacy, due to Dwork and McSherry (see [22]), that "no matter what an attacker knows ahead of time, the attacker learns the same information about any individual i from the mechanism whether or not that individual's data were used." It accomplishes the same primary goal as semantic privacy (Definition 3.4), and analogously to Theorem 3.1 it is equivalent to Definitions 3.3 and 5.1

5.3 A Distributional Version of Differential Privacy

As a warm-up to our general framework, we first describe a particular instantiation that we dub *distributional differential privacy* (DDP). The main idea is that, rather than require indistinguishability to hold for *all* distributions over the dataset, we require it to hold only for some specified set of "candidate" distributions Δ. (One

[1] We note that this informally stated theorem uses ϵ, δ-indistinguishability as a measure of closeness between distributions, whereas we defined semantic privacy (Definition 3.4) using statistical distance. Either can be used with analogous results (with different parameters) depending on preference.

can view the set of candidate distributions for X as representing the possibilities for the "true" distribution of the data, or as representing the adversary's possible uncertainty about the data.) We present two variants of the definition. The first, which we view as more intuitively appealing, is obtained by relaxing the inference-based definition discussed in the previous section. The second, which can be viewed as a relaxation of the simulation-based definition of Gehrke et al. [30], is somewhat easier to work with and, importantly, is strictly stronger than our inference-based formulation.

We obtain a distributional variant of the inference-based definition from the previous section by requiring Equation (5.1) to hold only for some set of candidate distributions Δ over X, rather than for all possible distributions. Fix some class Δ of probability distributions over random variables $(X, Z) \in \mathcal{U}^* \times \{0,1\}^*$, where X represents the dataset and Z denotes auxiliary information known to the adversary. We then have:

Definition 5.2 *A mechanism F satisfies* $(\epsilon, \delta, \Delta)$-inference-based distributional differential privacy *if there is a simulator* Sim *such that for all distributions $\mathcal{D} \in \Delta$ on (X, Z), with probability at least $1 - \delta$ over choice of $(t, z) = (F(X), Z)$ the following holds for all i:*

$$X_i\big|_{F(X)=t,\, Z=z} \approx_{\epsilon,\delta} X_i\big|_{\mathsf{Sim}(X_{-i})=t,\, Z=z}\,. \tag{5.2}$$

A variant is obtained by generalizing the simulation-based definition of Gehrke et al. [30].

Definition 5.3 *A mechanism F satisfies* $(\epsilon, \delta, \Delta)$-distributional differential privacy *if*

there is a simulator Sim *such that for all distributions* $\mathcal{D} \in \Delta$ *on* (X, Z)*, all* i*, and all* $(x_i, z) \in \mathsf{Supp}(X_i, Z)$*:*

$$F(X)\big|_{X_i=x_i,\, Z=z} \approx_{\epsilon,\delta} \mathsf{Sim}(X_{-i})\big|_{X_i=x_i,\, Z=z}\,. \tag{5.3}$$

In Chapter 6 we will show several example DDP mechanisms. In both Definition 5.2 and 5.3 taking Δ to be the set of all distributions (or simply all point distributions) gives differential privacy. However, in general DDP is stronger than inference-based DDP.

Theorem 5.1 *Say* F *satisfies* $(\epsilon, \delta, \Delta)$*-DDP where distributions in* Δ *have support only on datasets of size at most* n*, and* $2\sqrt{\delta n} \leq \epsilon e^{\epsilon}$*. Then* F *satisfies* $(3\epsilon, 2\sqrt{\delta n}, \Delta)$*-inference-based DDP.*

Theorem 5.1 is a special case of Theorem 5.2, which we prove in the next section. Theorems 5.1 and 5.2 are both generalizations of a result of [45], who proved the same statement for the usual notion of differential privacy.

The converse of Theorem 5.1 does not hold in general, as the following example shows.

Example 5.1 (Inference-based vs. indistinguishability-based DDP) *Let* Δ *contain a single distribution on* (X, Z)*, where* $X = (X_1, \ldots, X_n)$ *is a tuple of* n *uniformly distributed bits and* $Z = \oplus_{i=1}^{n} X_i$*. Say* $F(X)$ *outputs the parity of its input. Note that for any* $x_i, z \in \{0, 1\}$*, the distribution* $F(X)\big|_{X_i=x_i, Z=z}$ *is just a point distribution on the value* z*. However,* X_{-i} *(and hence* $\mathsf{Sim}(X_{-i})$*) is independent of*

$F(X) = Z$, and so the distribution of $\mathsf{Sim}(X_{-i})$ cannot equal Z with probability better than 1/2. Thus, conditioned on Z, the distributions of $F(X)$ and $\mathsf{Sim}(X_{-i})$ are very different in general, and so F cannot satisfy DDP for any reasonable parameters.

On the other hand, for any t, z the distribution $X_i\big|_{F(X)=t,Z=z}$ is uniform. If Sim outputs a uniform bit, then $X_i\big|_{F(X)=t,Z=z} = X_i\big|_{\mathsf{Sim}(X_{-i})=t,Z=z}$ and so F does satisfy inference-based DDP.

5.3.1 General Framework

Distributional differential privacy is just one possible instantiation of a general framework we call *coupled-worlds (CW) privacy*. At a high level, definitions within our framework are specified by two functions[2] alt and priv; if a mechanism F satisfies the definition then, intuitively, "$F(X)$ reveals no more information about $\mathsf{priv}(X)$ than is revealed by $\mathsf{alt}(X)$." That is, priv allows one to specify what information should be kept private, while alt defines a "scrubbed" version of the dataset that is available in some ideal world. For the specific case of DDP, we are interested in the privacy of an individual record X_i (so $\mathsf{priv}(X) = X_i$), and want to ensure that $F(X)$ reveals no more information about X_i than would be revealed if user i had not participated in the study at all (so $\mathsf{alt}(X) = X_{-i}$).

We start with an inference-based version of our framework that we find intuitively compelling.

Definition 5.4 *A mechanism F satisfies $(\epsilon, \delta, \Delta, \Gamma)$-*inference-based coupled-worlds privacy *if there is a simulator Sim such that for all distributions $\mathcal{D} \in \Delta$ on (X, Z),*

[2]Formally, they are specified by a *set* $\Gamma = \{(\mathsf{alt}_i, \mathsf{priv}_i)\}$ of function pairs.

with probability at least $1-\delta$ *over choice of* $(t,z) = (F(X), Z)$ *the following holds for all* $(\mathsf{alt}, \mathsf{priv}) \in \Gamma$*:*

$$\mathsf{priv}(X)\big|_{F(X)=t,\, Z=z} \approx_{\epsilon,\delta} \mathsf{priv}(X)\big|_{\mathsf{Sim}(\mathsf{alt}(X))=t,\, Z=z}. \tag{5.4}$$

As with DDP, it is convenient to use an alternate, indistinguishability-based definition which implies the inference-based version.

Definition 5.5 *A mechanism* F *satisfies* $(\epsilon, \delta, \Delta, \Gamma)$-coupled worlds privacy *if there is a simulator* Sim *such that for all distributions* $\mathcal{D} \in \Delta$ *on* (X, Z)*, all* $(\mathsf{alt}, \mathsf{priv}) \in \Gamma$*, and all* $(v, z) \in \mathsf{Supp}(\mathsf{priv}(X), Z)$*:*

$$F(X)\big|_{\mathsf{priv}(X)=v,\, Z=z} \approx_{\epsilon,\delta} \mathsf{Sim}(\mathsf{alt}(X))\big|_{\mathsf{priv}(X)=v,\, Z=z}.$$

Theorem 5.2 *Say* F *satisfies* $(\epsilon, \delta, \Delta, \Gamma)$*-CW privacy, where* $2\sqrt{\delta|\Gamma|} \leq \epsilon e^{\epsilon}$*. Then* F *satisfies* $(3\epsilon, 2\sqrt{\delta|\Gamma|}, \Delta, \Gamma)$*-inference-based CW privacy.*

The proof of Theorem 5.2 relies on a generalization of [45, Lemma 4.1], as follows:

Lemma 5.1 *Suppose* $(A, B) \approx_{\epsilon,\delta} (A', B')$*. Then, for every* $\delta_2 > 0$ *and* $\delta_1 = \frac{2\delta}{\delta_2} + \frac{2\delta}{\epsilon e^{\epsilon}}$*, the following holds: with probability at least* $1-\delta_1$ *over* t *chosen according to* B*, the random variables* $A|_{B=t}$ *and* $A'|_{B'=t}$ *are* $(3\epsilon, \delta_2)$*-indistinguishable.*

Proof of Theorem 5.2: Fix a mechanism F with simulator Sim, a distribution $\mathcal{D}$

in Δ, and a pair $(\mathsf{alt}, \mathsf{priv}) \in \Gamma$. CW privacy implies that:

$$(F(X), \mathsf{priv}(X), Z) \approx_{\epsilon,\delta} (\mathsf{Sim}(\mathsf{alt}(X)), \mathsf{priv}(X), Z)\,.$$

Take $\delta_2 = 2\sqrt{\delta|\Gamma|}$ and $\delta_1 = \frac{2\delta}{\delta_2} + \frac{2\delta}{\epsilon e^\epsilon}$. We can apply Lemma 5.1 with $A = A' = \mathsf{priv}(X)$, $B = (F(X), Z)$, and $B' = (\mathsf{Sim}(\mathsf{alt}(X)), Z)$ to get that with probability $1 - \delta_1$ over (t, z), we have

$$\mathsf{priv}(X)\big|_{F(X)=t,Z=z} \approx_{3\epsilon,\delta_2} \mathsf{priv}(X)\big|_{\mathsf{Sim}(\mathsf{alt}(X))=t,Z=z}.$$

Taking a union bound over all function pairs in Γ, we see that the above holds for all $(\mathsf{alt}, \mathsf{priv}) \in \Gamma$ with probability at least

$$1 - |\Gamma| \cdot \delta_1 = 1 - |\Gamma| \cdot \left(\frac{2\delta}{2\sqrt{\delta|\Gamma|}} + \frac{2\delta}{\epsilon e^\epsilon}\right) = 1 - \sqrt{\delta|\Gamma|} - \frac{2\delta|\Gamma|}{\epsilon e^\epsilon} \geq 1 - 2\sqrt{\delta|\Gamma|}\,,$$

where the final inequality follows because $\epsilon e^\epsilon \geq 2\sqrt{\delta|\Gamma|}$. □

As noted earlier for the specific case of DDP, the implication in Theorem 5.2 is strict.

Other instantiations. We have already discussed one instantiation of CW privacy (namely, distributional differential privacy) in the previous section. We briefly mention some other interesting instantiations.

- Consider a database representing a social network. Here, the database is a graph and private data is associated with each node or edge. We can define a

version of node-level privacy by taking pairs $(\mathsf{alt}, \mathsf{priv})$ in which priv outputs information associated with a given node and its incident edges, and alt removes that node and its incident edges.

- Frequently some data (say, demographic information like sex and age) is public and need not be protected. To model this we can consider pairs $(\mathsf{alt}_i, \mathsf{priv}_i)$ in which priv_i outputs only the private data in record X_i and alt_i removes only the private information.

In all the examples we have discussed so far, alt and priv are complementary. This need not always be the case:

- Imagine a database in which several schools contribute data of their students. In this situation each school might want to make sure that no more can be learned about each of its students than if the *entire school* had chosen not to participate in the study. To model this we can consider pairs $(\mathsf{alt}, \mathsf{priv})$ in which priv still outputs an individual student's record, but alt removes all records associated with that student's school.

- Suppose a study involves a database of assets of several financial firms. Having alt remove all the data of any single firm might be too limiting. Instead we might only require that a certain amount of ambiguity about each firm's data remains. This could be achieved by letting alt add noise to the asset distribution of a firm.

5.4 Properties of the Framework

We now explore several properties of the CW privacy framework. These properties serve two distinct purposes. The first is to further confirm that the definition is consistent with our intuitive understanding of privacy. If the abstract idea of privacy obeys certain properties, then a definition attempting to formalize that idea should obey those properties as well. The second purpose is to provide tools that are useful for the development and analysis of private mechanisms.

We first show that CW privacy is preserved under post-processing. This is basically a proof that the definition is not a superficial property of the output's formatting, but rather a true limitation on its actual information content. Any reasonable privacy definition must satisfy this property.

Theorem 5.3 *Coupled-worlds privacy is preserved under post-processing. Formally, if mechanism F satisfies $(\epsilon, \delta, \Delta, \Gamma)$-CW privacy, then so does $G \circ F$ for any (randomized) function G.*

Proof: Suppose that F is $(\epsilon, \delta, \Delta, \Gamma)$-CW private. Let Sim be the required simulator for F, and fix any distribution $\mathcal{D} \in \Delta$ on (X, Z), any $(\mathsf{alt}, \mathsf{priv}) \in \Gamma$, and any v, z in the support of $(\mathsf{priv}(X), Z)$. We also say that the randomness of G is achieved through the use of random coins, and fix a series γ of such coins. Now that γ is fixed, G is a deterministic function and we let $G^{-1}(S)$ denote the pre-image of S

under this function. For any set S, observe that

$$\begin{aligned}
&\Pr\left[G(F(X)) \in S \mid \mathsf{priv}(X) = v, Z = z\right] \\
&= \Pr\left[F(X) \in G^{-1}(S) \mid \mathsf{priv}(X) = v, Z = z\right] \\
&\leq e^{\epsilon} \cdot \Pr\left[\mathsf{Sim}(\mathsf{alt}(X)) \in G^{-1}(S) \mid \mathsf{priv}(X) = v, Z = z\right] + \delta \\
&= e^{\epsilon} \cdot \Pr\left[G\left(\mathsf{Sim}(\mathsf{alt}(X))\right) \in S \mid \mathsf{priv}(X) = v, Z = z\right] + \delta
\end{aligned}$$

Since this inequality holds for each choice of randomness γ, it also holds when the two probabilities are each averaged over all possible choices of γ. The other direction is the same as above, proving that

$$G(F(X))\big|_{\mathsf{priv}(X)=v,\, Z=z} \approx_{\epsilon,\delta} G(\mathsf{Sim}(\mathsf{alt}(X)))\big|_{\mathsf{priv}(X)=v,\, Z=z}\,.$$

Hence, $G(F(\cdot))$ is $(\epsilon, \delta, \Delta, \Gamma)$-CW private. Note that the simulator for $G(F(\cdot))$ is $G(\mathsf{Sim}(\cdot))$. □

The next two results show that if a mechanism F satisfies CW privacy with respect to some class of distributions Δ, then it also satisfies CW privacy with respect to a (potentially) larger class Δ'. In the first case, we show that one can take Δ' to include all distributions that are convex combinations of distributions in Δ. Besides being a desirable property in its own right, it also serves as a technically convenient tool and is used heavily in our analysis of particular mechanisms. First we formally define convex combinations.

Definition 5.6 *A* convex combination $\mathcal{D}'$ *of distributions in* Δ *is a distribution*

achieved by attaching a weight $\lambda_{\mathcal{D}}$ *to each distribution* $\mathcal{D} \in \Delta$ *with* $\sum_{\mathcal{D}\in\Delta} \lambda_{\mathcal{D}} = 1$ *and setting*

$$\Pr[D = c \mid D \leftarrow \mathcal{D}'] = \sum_{\mathcal{D}\in\Delta} \lambda_{\mathcal{D}} \Pr[D = c \mid D \leftarrow \mathcal{D}]$$

for each possible output c*. If the distributions are continuous, replace the probability of outputting* c *with the probability density function's value at* c*. The* convex hull *of* Δ *is the set of all possible convex combinations of the elements of* Δ*.*

We now present the theorem.

Theorem 5.4 *If* F *satisfies* $(\epsilon, \delta, \Delta, \Gamma)$*-CW privacy, it also satisfies* $(\epsilon, \delta, \Delta', \Gamma)$*-CW privacy for* Δ' *the convex hull of* Δ*. That is,* Δ' *is the set of all convex combinations of distributions in* Δ*.*

Proof: Let Sim be the required simulator for F. Let $\mathcal{D}_3 \in \Delta'$, be a convex combination of two distributions $\mathcal{D}_1, D_2 \in \Delta$ where $\mathcal{D}_1$ and $\mathcal{D}_2$ are sampled with probabilities λ and $1-\lambda$ respectively. (We deal with two distributions for simplicity, but the proof for combinations of more follows along the same lines.) We use $\Pr_1$ (resp. $\Pr_2, \Pr_3$) to denote a probability over (X, Z) drawn from $\mathcal{D}_1$ (resp. $\mathcal{D}_2, \mathcal{D}_3$). Fix any $(\mathsf{alt}, \mathsf{priv}) \in \Gamma$, and (v, z) in the support of $(\mathsf{priv}(X), Z)$ (when (X, Z) is

drawn from $\mathcal{D}_3$). Then for any S:

$$
\begin{aligned}
&\Pr_3[F(X) \in S, \mathsf{priv}(X) = v, Z = z] \\
&= \lambda \cdot \Pr_1[F(X) \in S, \mathsf{priv}(X) = v, Z = z] \\
&\quad +(1-\lambda) \cdot \Pr_2[F(X) \in S, \mathsf{priv}(X) = v, Z = z] \\
&\leq e^{\epsilon} \cdot \Big(\lambda \cdot \Pr_1[\mathsf{Sim}(\mathsf{alt}(X)) \in S, \mathsf{priv}(X) = v, Z = z] \\
&\qquad + (1-\lambda) \cdot \Pr_2[\mathsf{Sim}(\mathsf{alt}(X)) \in S, \mathsf{priv}(X) = v, Z = z]\Big) \\
&\quad + \delta \cdot \lambda \cdot \Pr_1[\mathsf{priv}(X) = v, Z = z] + \delta \cdot (1-\lambda) \cdot \Pr_2[\mathsf{priv}(X) = v, Z = z] \\
&\leq e^{\epsilon} \cdot \Pr_3[\mathsf{Sim}(\mathsf{alt}(X)) \in S, \mathsf{priv}(X) = v, Z = z] \\
&\quad + \delta \cdot \lambda \cdot \Pr_1[\mathsf{priv}(X) = v, Z = z] + \delta \cdot (1-\lambda) \cdot \Pr_2[\mathsf{priv}(X) = v, Z = z] \\
&\leq e^{\epsilon} \cdot \Pr_3[\mathsf{Sim}(\mathsf{alt}(X)) \in S, \mathsf{priv}(X) = v, Z = z] + \delta \cdot \Pr_3[\mathsf{priv}(X) = v, Z = z]
\end{aligned}
$$

where the subscript indicates the distribution under consideration. Hence,

$$
\begin{aligned}
&\Pr_3[F(X) \in S \mid \mathsf{priv}(X) = v, Z = z] \\
&\quad = \frac{\Pr_3[F(X) \in S, \mathsf{priv}(X) = v, Z = z]}{\Pr_3[\mathsf{priv}(X) = v, Z = z]} \\
&\quad \leq \frac{e^{\epsilon} \cdot \Pr_3[\mathsf{Sim}(\mathsf{alt}(X)) \in S, \mathsf{priv}(X) = v, Z = z] + \delta \cdot \Pr_3[\mathsf{priv}(X) = v, Z = z]}{\Pr_3[\mathsf{priv}(X) = v, Z = z]} \\
&\quad \leq e^{\epsilon} \cdot \Pr_3[\mathsf{Sim}(\mathsf{alt}(X)) \in S \mid \mathsf{priv}(X) = v, Z = z] + \delta.
\end{aligned}
$$

The other direction of the inequality is proved in the same way. □

Next, we show that CW privacy continues to hold if the attacker's auxiliary information is reduced. Again, besides being a desirable property in its own right,

it is also technically useful since it then suffices to prove CW privacy of some mechanism only with respect to some realistic *upper bound* on the auxiliary information available to an adversary.

Theorem 5.5 *If F satisfies $(\epsilon, \delta, \Delta, \Gamma)$-CW privacy, then it satisfies $(\epsilon, \delta, \Delta', \Gamma)$-CW privacy for Δ', where $\mathcal{D}' \in \Delta'$ first samples (X, Z) from some $\mathcal{D} \in \Delta$, then outputs (X, Z') with $Z' = f(Z)$ for some (randomized) function f.*

Proof: Suppose that F is $(\epsilon, \delta, \Delta, \Gamma)$-CW private. Let Sim be the simulator for F. Fix a distribution $\mathcal{D} \in \Delta$ on (X, Z) and let Z' be as described in the theorem statement. Clearly, $X \to Z \to Z'$ is a Markov chain, i.e., Z can be viewed as a randomized function of X, and Z' as a randomized function of Z. Note that this also implies that $(X, F(X), \mathsf{Sim}(\mathsf{alt}(X)), \mathsf{priv}(X)) \to Z \to Z'$ is a Markov chain. Fix $(\mathsf{priv}, \mathsf{alt}) \in \Gamma$, $S \subseteq \mathrm{Range}(F(X))$, and $(v, z') \in \mathsf{Supp}(\mathsf{priv}(X), Z')$. Let $T_Z(v, z') = \mathsf{Supp}(Z|_{\mathsf{priv}(X)=v, Z'=z'})$. Observe that

$$
\begin{aligned}
&\Pr[F(X) \in S \mid \mathsf{priv}(X) = v, Z' = z'] \\
&= \sum_{z \in T_Z(v,z')} \Pr[F(X) \in S \mid \mathsf{priv}(X) = v, Z = z, Z' = z'] \\
&\qquad \times \Pr[Z = z \mid \mathsf{priv}(X) = v, Z' = z'] \\
&= \sum_{z \in T_Z(v,z')} \Pr[F(X) \in S \mid \mathsf{priv}(X) = v, Z = z] \\
&\qquad \times \Pr[Z = z \mid \mathsf{priv}(X) = v, Z' = z'] \qquad (5.5) \\
&\leq \sum_{z \in T_Z(v,z')} (e^{\epsilon} \Pr[\mathsf{Sim}(\mathsf{alt}(X)) \in S \mid \mathsf{priv}(X) = v, Z = z] + \delta) \\
&\qquad \times \Pr[Z = z \mid \mathsf{priv}(X) = v, Z' = z']
\end{aligned}
$$

$$= e^{\epsilon} \sum_{z \in T_Z(v,z')} \Pr\left[\mathsf{Sim}\left(\mathsf{alt}(X)\right) \in S \mid \mathsf{priv}(X) = v, Z = z, Z' = z'\right]$$

$$\times \Pr[Z = z \mid \mathsf{priv}(X) = v, Z' = z'] + \delta \tag{5.6}$$

$$= e^{\epsilon} \Pr\left[\mathsf{Sim}\left(\mathsf{alt}(X)\right) \in S \mid \mathsf{priv}(X) = v, Z' = z'\right] + \delta$$

where (5.5) and (5.6) follow from the fact that $(X, F(X), \mathsf{Sim}\left(\mathsf{alt}(X)\right), \mathsf{priv}(X)) \to Z \to Z'$ is a Markov chain. (In particular, in a Markov chain $A \to B \to C$, A and C are independent conditioned on B. In this case that means that conditioning on $Z' = z$ does not affect the probability in question.) □

Finally, we turn to the question of the *composition* of two private mechanisms F and G. Here, both $F(X)$ and $G(X)$ are released. Although we are not able to prove as general a composition theorem as we would like, we can show that the composition satisfies CW privacy as long as G is private even when given $F(X)$ as auxiliary information, and F is private when given $\mathsf{Sim}_G(\mathsf{alt}(X))$ as auxiliary information.

Theorem 5.6 *Let F and G be two mechanisms, Δ a class of distributions, and Γ a family of $(\mathsf{priv}, \mathsf{alt})$ pairs. Say G is $(\epsilon_G, \delta_G, \Delta_G, \Gamma)$-CW private with simulator Sim_G, where Δ_G includes all distributions that output $(X, (Z, F(X)))$ where (X, Z) is drawn from some $\mathcal{D} \in \Delta$. Say F is $(\epsilon_F, \delta_F, \Delta_F, \Gamma)$-CW private with simulator Sim_F, where Δ_F includes all distributions that output $\left(X, \left(Z, \mathsf{Sim}_G\left(\mathsf{alt}(X)\right)\right)\right)$ where (X, Z) is drawn from some $\mathcal{D} \in \Delta$ and alt is the first element of a pair in Γ. Then*

the mechanism $H = (F, G)$ *is* $(\epsilon_H, \delta_H, \Delta, \Gamma)$*-CW private where*

$$\epsilon_H = \epsilon_F + \epsilon_G$$

$$\delta_H = \max\left(\delta_F e^{\epsilon_G} + \delta_G,\ \delta_F + \delta_G e^{\epsilon_F}\right) = O(\delta_F + \delta_G).$$

Proof: We will show that the mechanism H is $(\epsilon_H, \delta_H, \Delta, \Gamma)$-CW private with respect to a simulator Sim_H defined as $(\mathsf{Sim}_F, \mathsf{Sim}_G)$. Let S be any subset of $\mathsf{Supp}\left(F(X)\right) \times \mathsf{Supp}\left(G(X)\right)$. Let S_1 be the set of all first-elements of the pairs in S and $S_2(s_1)$ be the set of all values s_2 such that $(s_1, s_2) \in S$. Let priv be any private information function given by the family Γ. Let v be any element in the support of $\mathsf{priv}(X)$ and z be any element in the support of the auxiliary information Z. Note that we can write

$$\begin{aligned}
&\Pr\left[H(X) \in S | \mathsf{priv}(X) = v, Z = z\right] \\
&= \sum_{s_1 \in \S_1} \Pr\left[G(X) \in S_2(s_1) | F(X) = s_1, \mathsf{priv}(X) = v, Z = z\right] \\
&\qquad \times \Pr\left[F(X) = s_1 | \mathsf{priv}(X) = v, Z = z\right]
\end{aligned}$$

Since conditioning on $\mathsf{priv}(X) = v,\ Z = z$ will be in every probability term in this proof, we will drop such conditioning just to keep the notation manageable, but such conditioning should be implicitly understood.

Now, observe that

$$
\begin{aligned}
&\Pr[H(X) \in S] \\
&= \sum_{s_1 \in S_1} \Pr[G(X) \in S_2(s_1) | F(X) = s_1] \Pr[F(X) = s_1] \\
&= \sum_{s_1 \in S_1} \Bigg(e^{\epsilon_G} \Pr[\mathsf{Sim}_G(\mathsf{alt}(X)) \in S_2(s_1) | F(X) = s_1] \\
&\qquad + \Pr[G(X) \in S_2(s_1) | F(X) = s_1] \\
&\qquad - e^{\epsilon_G} \Pr[\mathsf{Sim}_G(\mathsf{alt}(X)) \in S_2(s_1) | F(X) = s_1] \Bigg) \Pr[F(X) = s_1] \\
&\leq e^{\epsilon_G} \Pr[(\mathsf{Sim}_G(\mathsf{alt}(X)),\ F(X)) \in S] + \delta_G \qquad (5.7) \\
&= e^{\epsilon_G} \sum_{s_1 \in S_1} \sum_{s_2 \in S_2(s_1)} \Pr[F(X) = s_1 | \mathsf{Sim}_G(\mathsf{alt}(X)) = s_2] \Pr[\mathsf{Sim}_G(\mathsf{alt}(X)) = s_2] + \delta_G \\
&\leq e^{\epsilon_G} \sum_{s_1 \in S_1} \sum_{s_2 \in S_2(s_1)} (e^{\epsilon_F} \Pr[\mathsf{Sim}_F(\mathsf{alt}(X)) = s_1 | \mathsf{Sim}_G(\mathsf{alt}(X)) = s_2] + \delta_F) \\
&\qquad \times \Pr[\mathsf{Sim}_G(\mathsf{alt}(X)) = s_2] + \delta_G \qquad (5.8) \\
&\leq e^{\epsilon_F + \epsilon_G} \Pr[(\mathsf{Sim}_F(\mathsf{alt}(X)),\ \mathsf{Sim}_G(\mathsf{alt}(X))) \in S] + e^{\epsilon_G} \delta_F + \delta_G \\
&= e^{\epsilon_F + \epsilon_G} \Pr[\mathsf{Sim}_H(\mathsf{alt}(X)) \in S] + e^{\epsilon_G} \delta_F + \delta_G
\end{aligned}
$$

where (5.7) follows from the fact that G is $(\epsilon_G, \delta_G, \Delta_G, \Gamma)$-CW private with respect to Sim_G and (5.8) follows from the fact that F is $(\epsilon_F, \delta_F, \Delta_F, \Gamma)$-CW private with respect to Sim_F.

Similarly, one can show that

$$\Pr[\mathsf{Sim}_H(\mathsf{alt}(X)) \in S] \leq e^{\epsilon_F + \epsilon_G} \Pr[H(X) \in S] + e^{\epsilon_F} \delta_G + \delta_F$$

Hence, the proof is complete. □

5.5 Relation to Other Definitions

We conclude our definitional treatment by comparing our definition to two other recent proposals: noiseless privacy [6] and Pufferfish [47].

Noiseless privacy was introduced with a similar motivation as our own; the idea was to use adversarial uncertainty about the dataset to eliminate the need for noise in the mechanism itself. The high-level idea is to require that $F(X)$ "looks similar" for any two values of a given record:

Definition 5.7 *F satisfies $(\epsilon, \delta, \mathcal{D})$-*noiseless privacy *if for all i, z, and neighboring x_i and x_i':*

$$F(X)\big|_{X_i=x_i, Z=z} \approx_{\epsilon,\delta} F(X)\big|_{X_i=x_i', Z=z},$$

where (X, Z) is chosen according to distribution $\mathcal{D}$.

When $\delta > 0$ this definition is slightly different from the version in [6]. In particular, we require (ϵ, δ)-indistinguishability to hold for all choices of x_i and x_i', whereas the definition in [6] requires ϵ-indistinguishability to hold except for x_i, x_i' that occur with probability at most δ.

Pufferfish provides a framework for defining privacy. Noiseless privacy can be viewed as one specific instantiation,[3] but others are possible. Pufferfish allows for customization of what information will be kept private by appropriate choice of a

[3]This is true for the definitions as given here, which differ slightly from the definitions given in the original works.

function sec, which takes as input a dataset X and outputs an element of $\{0, 1, \perp\}$. Thus, sec defines two disjoint classes of datasets, the preimages of 0 and 1, with the $\perp$ output allowing the function to be indecisive. Roughly, Pufferfish defines a mechanism F to be private if the distribution of $F(X)$ is similar regardless of which value of $\mathsf{sec}(X)$ we condition on.

Definition 5.8 *A mechanism F satisfies $(\epsilon, \delta, \Delta, \mathcal{S})$-*Pufferfish privacy *if for all $\mathsf{sec} \in \mathcal{S}$, all z, and all distributions (X, Z) in Δ it holds that:*

$$F(X)\big|_{\mathsf{sec}(X)=0,\ Z=z} \approx_{\epsilon,\delta} F(X)\big|_{\mathsf{sec}(X)=1,\ Z=z} \tag{5.9}$$

(This definition differs in some non-essential ways from the definition in [47]. We highlight that we allow $\delta > 0$, something not done in [47].)

In both noiseless privacy (and, by extension, Pufferfish) and our notion of distributional definition privacy, the requirement is that $F(X)$ should be "roughly the same" in each of two possible worlds. The difference between the definitions is in which two worlds are compared. In noiseless privacy and Pufferfish the comparison is between a world in which X_i (resp., $\mathsf{sec}(X)$) takes on one value and a world in which it takes on some other value. In DDP, in contrast, the comparison is between a world in which X_i is included in the dataset and a world in which it is not. This has significant implications. Consider an example in which there is a global parameter μ which is either $+1$ or -1 (with half probability each), and every record is normally distributed with mean μ and standard deviation much smaller than 1. Note that the records are dependent because they all depend on the value

of μ. (They are, however, independent conditioned on μ.) The mechanism F that computes the sample mean $\bar{X}$ of the dataset and then outputs ± 1 depending on which is closer to $\bar{X}$ does *not* satisfy noiseless privacy: the distribution of $F(X)$ conditioned on $X_i \approx -1$ is very different from the distribution of $F(X)$ conditioned on $X_i \approx +1$. On the other hand, F *does* satisfy DDP (with the obvious simulator that simply runs F) since the distributions of $F(X)$ and $F(X_{-i})$ are close for X sampled according to the stated distribution.

To see the difference between our definitions and prior ones, it may help to consider an inference-based version of Pufferfish.

Definition 5.9 *A mechanism F satisfies $(\epsilon,\ \delta,\ \Delta,\ \mathcal{S})$-*inference-based Pufferfish privacy *if for all $\mathsf{sec} \in \mathcal{S}$, all z, and all distributions (X, Z) in Δ, with probability $1-\delta$ over choice of $t \leftarrow F(X)|_{Z=z}$ we have*

$$\mathsf{sec}(X)|_{F(X)=t,Z=z} \approx_{\epsilon,\delta} \mathsf{sec}(X)|_{Z=z}.$$

We have the following theorem, which shows that the inference-based version of Pufferfish privacy is implied by the standard version, as long as sec never outputs $\perp$. A similar version was proved in [47] (for the $\delta = 0$ case).

Theorem 5.7 *Say F satisfies $(\epsilon, \delta, \Delta, \mathcal{S})$-Pufferfish privacy, where all $\mathsf{sec} \in S$ have output in $\{0, 1\}$ and $2\sqrt{\delta} < \epsilon e^{\epsilon}$. Then it also satisfies $(3\epsilon, 2\sqrt{\delta}, \Delta, \mathcal{S})$-inference-based Pufferfish privacy.*

Proof: Because F is Pufferfish private, we know that for all sets S

$$\Pr[F(X) \in S \mid \mathsf{sec}(X) = 0, Z = z] \leq e^{\epsilon} \Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] + \delta \quad (5.10)$$
$$\Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] \leq e^{\epsilon} \Pr[F(X) \in S \mid \mathsf{sec}(X) = 0, Z = z] + \delta$$

Now, we can write $\Pr[F(X) \in S]$ as the sum of such a probability conditioned on the two possible values of $\mathsf{sec}(X)$.

$$\begin{aligned}
&\Pr[F(X) \in S \mid Z = z] \\
&\quad = \Pr[F(X) \in S \mid \mathsf{sec}(X) = 0, Z = z] \Pr[\mathsf{sec}(X) = 0] \\
&\qquad + \Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] \Pr[\mathsf{sec}(X) = 1] \\
&\quad \leq (e^{\epsilon} \Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] + \delta) \Pr[\mathsf{sec}(X) = 0] \\
&\qquad + \Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] \Pr[\mathsf{sec}(X) = 1]
\end{aligned}$$

Because $\Pr[\mathsf{sec}(X) = 1] + \Pr[\mathsf{sec}(X) = 0] = 1$ this last line is a weighted average of two values, and we can upper bound it by the greater of the two values to get

$$\Pr[F(X) \in S \mid Z = z] \leq e^{\epsilon} \Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] + \delta \quad (5.11)$$

We now need a similar bound in the opposite direction. We use equation 5.11 to obtain

$$\Pr[F(X) \in S \mid \mathsf{sec}(X) = 0, Z = z] \geq e^{-\epsilon}(\Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] - \delta)$$

We then follow similar computations to those above.

$$\begin{aligned}\Pr[F(X) \in S \mid Z = z] = &\Pr[F(X) \in S \mid \mathsf{sec}(X) = 0, Z = z] \Pr[\mathsf{sec}(X) = 0] \\ &+ \Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] \Pr[\mathsf{sec}(X) = 1] \\ \geq &e^{-\epsilon}(\Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] - \delta) \Pr[\mathsf{sec}(X) = 0] \\ &+ \Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] \Pr[\mathsf{sec}(X) = 1]\end{aligned}$$

Again, this last line is a weighted average of two values, and we can lower bound it by the lesser of the two values to get

$$\begin{aligned}\Pr[F(X) \in S \mid Z = z] &\geq e^{-\epsilon}(\Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] - \delta) \\ \Pr[F(X) \in S \mid \mathsf{sec}(X) = 1, Z = z] &\leq e^{\epsilon} \Pr[F(X) \in S \mid Z = z] + \delta \qquad (5.12)\end{aligned}$$

Using equations 5.11 and 5.12 (and the fact that the $\mathsf{sec}(X) = 0$ and $\mathsf{sec}(X) = 1$ cases are completely symmetric) we get that for any secret value s we have

$$F(X)|_{\mathsf{sec}(X)=s, Z=z} \approx_{\epsilon,\delta} F(X)|_{Z=z}$$

We now look at two joint distributions. The first is of $(F(X), \mathsf{sec}(X))$. We remove the explicit conditioning on $Z = z$, but this is only a change in notation. We assume that X is drawn from the conditional distribution given that $Z = z$. The second distribution is $(F(X), \mathsf{sec}(Y))$. Here Y is a separate random variable drawn from the same distribution as X (again assuming its associated auxilliary

information has value z). Our goal is to show that these two distributions are close to each other. Both have the same range. We can represent any set S in the range as two sets S_0 and S_1 such that $S_i = \{s \mid (s,i) \in S\}$.

$$
\begin{aligned}
&\Pr[(F(X), \mathsf{sec}(X)) \in S] \\
&\quad = \Pr[F(X) \in S_0 \mid \mathsf{sec}(X) = 0] \Pr[\mathsf{sec}(X) = 0] \\
&\qquad + \Pr[F(X) \in S_1 \mid \mathsf{sec}(X) = 1] \Pr[\mathsf{sec}(X) = 1] \\
&\quad \leq (e^\epsilon \Pr[F(X) \in S_0] + \delta) \Pr[\mathsf{sec}(X) = 0] \\
&\qquad + (e^\epsilon \Pr[F(X) \in S_1] + \delta) \Pr[\mathsf{sec}(X) = 1] \\
&\quad \leq e^\epsilon (\Pr[F(X) \in S_0] \Pr[\mathsf{sec}(X) = 0] + \Pr[F(X) \in S_1] \Pr[\mathsf{sec}(X) = 1]) + \delta \\
&\quad \leq e^\epsilon (\Pr[F(X) \in S_0] \Pr[\mathsf{sec}(Y) = 0] + \Pr[F(X) \in S_1] \Pr[\mathsf{sec}(Y) = 1]) + \delta \\
&\quad \leq e^\epsilon \Pr[(F(X), \mathsf{sec}(Y)) \in S] + \delta
\end{aligned}
$$

Very similar logic gives the other bound, meaning that we have (for all z)

$$(F(X), \mathsf{sec}(X)) \approx_{\epsilon,\delta} (F(X), \mathsf{sec}(Y))$$

Lemma 5.1 then tells us that for any $\delta_1 > 0$ the following holds: with probability $1 - \delta_1$ over $t \leftarrow F(X)$ the variables $\mathsf{sec}(X)|_{F(X)=t}$ and $\mathsf{sec}(Y)|_{F(X)=t}$ are $(3\epsilon, \delta_2)$-indistinguishable, where $\delta_1 = \frac{2\delta}{\delta_2} + \frac{2\delta}{\epsilon e^\epsilon}$.

We reintroduce the explicit conditioning of all variables on $Z = z$. We also remove the conditioning on $F(X) = t$ from the second distribution, since it is

independent from the variable in question. Having done so, we can also return to using X instead of Y as the random variable in that case. As a result, we have

$$\mathsf{sec}(X)|_{F(X)=t,Z=z} \approx_{3\epsilon,\delta_2} \mathsf{sec}(X)|_{Z=z}$$

with probability $\delta_1 = \frac{2\delta}{\delta_2} + \frac{2\delta}{\epsilon e^\epsilon} > 0$.

All that remains is to choose convenient parameters. We can choose $\delta_2 = 2\sqrt{\delta}$ and use the assumption that $2\sqrt{\delta} < \epsilon e^\epsilon$ to bound δ_1 by

$$\delta_1 = \frac{2\delta}{\delta_2} + \frac{2\delta}{\epsilon e^\epsilon} \geq \frac{2\delta}{2\sqrt{\delta}} + \frac{2\delta}{2\sqrt{\delta}} = 2\sqrt{\delta}.$$

This gives us the complete theorem.

□

Returning to Definition 5.9, one may interpret Pufferfish as requiring that the distribution of any sensitive information be roughly identical both before and after the release of $F(X)$.[4] This means that releasing estimates of general population parameters (for instance, whether smoking and cancer are correlated) is a privacy violation because it implies something about the information of any individual. In fact, Pufferfish considers the privacy of every individual to be violated in such a setting, even if their data is not used at all. In contrast, inference-based coupled-

[4]This might be surprising because, as we emphasize, this is a different type of interpretation from that of differential privacy, but differential privacy can be obtained as an instantiation of Pufferfish. This is because to instantiate differential privacy, one sets *participation in the database* as the secret information, rather than properties of the individual's information. The Pufferfish-style interpretation of whether participation is disclosed implies the differential privacy-style interpretation of whether information about an individual's information is disclosed. This is a subtle distinction, but it is crucial that this implication is particular to this particular instantiation, and similar implications do not follow from other instantiations.

worlds privacy (cf. Definition 5.4) only requires that the distribution of any private information be roughly identical whether F is computed over the entire dataset or over a "scrubbed" version of the dataset.

We note also that Pufferfish and noiseless privacy do not satisfy analogues of Theorem 5.5, which requires that privacy can only increase as auxiliary information is reduced. In particular, consider the motivating example from earlier, where outputting which of two possible values was closest to the sample mean was DDP, but did not satisfy noiseless privacy. If the auxiliary information discloses the true population mean (i.e., $Z = \mu$) the mechanism *is* private, but that privacy vanishes if the auxiliary information is reduced to nothing.

Chapter 6: DDP Mechanisms

In this chapter we present a variety of distributionally differentially private mechanisms. Our first result shows that a broad class of queries we call *stable* can be released exactly. In particular, this class includes computation of maximum a posterior probability (MAP) estimators, which are valuable statistical tools. Next, we look at histograms, which we show can be released exactly as long as low-count bins are suppressed. This result is proved under the assumption of a sampling distribution, which is an interesting distribution to consider because it is so realistic — many real world databases are indeed formed by sampling at random from a larger population.

Finally, we look at releasing exact sums of database values. We show a very general result, allowing sums to be computed under a wide variety of continuous distributions (though the precise nature of the distribution can greatly affect the values of ϵ and δ). We then show how this can be used to compute other, more complex functions. As an important example, we consider linear regression, which can be reduced to computing a series of sums.

The work in sections 6.1 and 6.2 originally appeared in FOCS 2013 [3], while the work in sections 6.3 and 6.4 is new here.

6.1 Stable Functions

We now consider deterministic functions that are "stable" under a particular (class of) distributions, by which we mean that the removal of one record has a low probability of changing the output (i.e., with high probability $F(X) = F(X_{-i})$). This property is sufficient to guarantee $(0, \delta, \Delta)$-DDP. Furthermore, if we require the existence of a non-zero lower bound on all conditional probabilities of the output of $F(X)$ given $X_i = x_i$ and $Z = z$ then the mechanism is $(\epsilon, 0, \Delta)$-DDP. Formally, this gives us two sufficient conditions to prove that a mechanism is DDP. We first consider the conditions guaranteeing $(0, \delta, \Delta)$-DDP.

Theorem 6.1 *Consider a deterministic database mechanism* $F : \mathcal{U}^n \to \mathcal{R}$ *and a class of distributions* Δ *for the pair* (X, Z)*. Suppose* $\exists\ \delta > 0$ *such that* $\forall \mathcal{D} \in \Delta,\ \forall i \in [n],\ \forall (x_i, z) \in Supp(X_i, Z)$*,*

$$\Pr\left[F(X) \neq F(X_{-i}) \mid X_i = x_i, Z = z\right] < \delta$$

Then, F *is* $(0,\ \delta,\ \Delta)$*-DDP.*

Proof: Without loss of generality fix a given $\mathcal{D} \in \Delta$. We use $\mathsf{Sim}(X_{-i} = F(X_{-i})$. For every $t \in \mathcal{R}$, every $i \in [n]$, and every $(x_i, z) \in \mathrm{Supp}(X_i, Z)$, we have

$$\begin{aligned}
&\Pr\left[F(X) = t \mid X_i = x_i, Z = z\right] \\
&\quad = \Pr\left[F(X) = t,\ F(X) = F(X_{-i}) \mid X_i = x_i, Z = z\right] \\
&\qquad + \Pr\left[F(X) = t \mid X_i = x_i, Z = z,\ F(X) \neq F(X_{-i})\right] \\
&\qquad\quad \times \Pr\left[F(X) \neq F(X_{-i}) \mid X_i = x_i, Z = z\right] \\
&\quad < \Pr\left[F(X_{-i}) = t \mid X_i = x_i, Z = z\right] + \delta
\end{aligned}$$

Similarly, we can show

$$\Pr\left[F(X_{-i}) = t \mid X_i = x_i, Z = z\right] \ < \Pr\left[F(X) = t \mid X_i = x_i, Z = z\right] + \delta$$

which completes the proof. □

We now move to the $(\epsilon, 0, \Delta)$-DDP case, which requires a slightly more complex assumption.

Theorem 6.2 *Consider a deterministic function $F : \mathcal{U}^n \to \mathcal{R}$ where $\mathcal{R}$ is a finite set that does not depend on n and a distribution class Δ_n for the pair (X, Z) that outputs databases X of size n. If there exist $c > 0$ and $\mu_n \in (0, c)$ such that, for all $\mathcal{D} \in \Delta_n$, all $i \in [n]$, all $t \in \mathcal{R}$, and all $(x_i, z) \in Supp(X_i, Z)$, the following conditions*

hold simultaneously

$$\Pr\left[F(X) = t \mid X_i = x_i, Z = z\right] \geq c \tag{6.1}$$

$$\Pr\left[F(X) \neq F(X_{-i}) \mid X_i = x_i, Z = z\right] \leq \mu_n \tag{6.2}$$

then, F is $(\epsilon_n,\ 0,\ \Delta_n)$-DDP where $\epsilon_n = \ln\left(\frac{c+\mu_n}{c-\mu_n}\right)$. Moreover, if the above holds for all n, c does not depend on n, and $\mu_n \to 0$, then $\epsilon_n \to 0$.

Proof: Suppose that conditions (6.1) and (6.2) hold. Let $i \in [n]$, $(x_i, z) \in \text{Supp}(X_i, Z)$, $t \in \mathcal{S}$. Let

$$\begin{aligned}
\text{Num}_n(i, x_i, z, t) &= \Pr[F(X) = t \mid X_i = x_i, Z = z] \\
\text{Den}_n(i, x_i, z, t) &= \Pr[F(X_{-i}) = t \mid X_i = x_i, Z = z] \\
&= \text{Num}_n(i, x_i, z, t) A_n(i, x_i, z, t) \\
&\quad + (1 - \text{Num}_n(i, x_i, z, t))\, B_n(i, x_i, z, t) \\
R(i, x_i, z, t) &= \frac{\text{Num}_n(i, x_i, z, t)}{\text{Den}_n(i, x_i, z, t)}
\end{aligned}$$

where

$$\begin{aligned}
A_n(i, x_i, z, t) &= \Pr\left[F(X_{-i}) = t \mid X_i = x_i, Z = z,\ F(X) = t\right] \\
B_n(i, x_i, z, t) &= \Pr\left[F(X_{-i}) = t \mid X_i = x_i, Z = z,\ F(X) \neq t\right]
\end{aligned}$$

First, we set a lower bound on $A_n(i, x_i, z, t)$ as follows:

$$\begin{aligned}
& \Pr\left[F(X_{-i}) \neq F(X) \mid X_i = x_i, Z = z\right] < \mu_n \\
\Leftrightarrow \quad & \textstyle\sum_{t \in \mathcal{S}} \mathrm{Num}_n(i, x_i, z, t) \Pr\left[F(X_{-i}) \neq t \mid F(X) = t,\ X_i = x_i, Z = z\right] < \mu_n \\
\Rightarrow \quad & \Pr\left[F(X_{-i}) \neq t \mid F(X) = t,\ X_i = x_i, Z = z\right] < \tfrac{\mu_n}{\mathrm{Num}_{n(i,x_i,z,t)}} \\
\Rightarrow \quad & A_n(i, x_i, z, t) > 1 - \tfrac{\mu_n}{\mathrm{Num}_{n(i,x_i,z,t)}} \qquad (6.3)
\end{aligned}$$

Next, we set an upper bound on $B_n(i, x_i, z, t)$:

$$\begin{aligned}
& \Pr\left[F(X_{-i}) \neq F(X) \mid X_i = x_i, Z = z\right] < \mu_n \\
\Rightarrow \quad & \Pr\left[F(X_{-i}) = t,\ F(X) \neq t \mid X_i = x_i, Z = z\right] < \mu_n \\
\Rightarrow \quad & B_n(i, x_i, z, t) < \tfrac{\mu_n}{1 - \mathrm{Num}_{n(i,x_i,z,t)}} \qquad (6.4)
\end{aligned}$$

Thus, from (6.3) and (6.4), it is easy to see that $R(i, x_i, z, t)$ is bounded from below and above as follows:

$$\frac{\mathrm{Num}_n(i, x_i, z, t)}{\mathrm{Num}_n(i, x_i, z, t) + \mu_n} \leq \quad R(i, x_i, z, t) \leq \frac{\mathrm{Num}_{n(i,x_i,z,t)}}{\mathrm{Num}_{n(i,x_i,z,t) - \mu_n}}$$

Hence,

$$
\begin{aligned}
&|\ln R(i, x_i, z, t)| \\
&\quad \leq \max\left(\ln\left(\frac{\mathrm{Num}_n(i, x_i, z, t)}{\mathrm{Num}_n(i, x_i, z, t) - \mu_n}\right), \ln\left(\frac{\mathrm{Num}_n(i, x_i, z, t) + \mu_n}{\mathrm{Num}_n(i, x_i, z, t)}\right)\right) && (6.5) \\
&\quad \leq \ln\left(\frac{\mathrm{Num}_n(i, x_i, z, t) + \mu_n}{\mathrm{Num}_n(i, x_i, z, t) - \mu_n}\right) && (6.6) \\
&\quad \leq \ln\left(\frac{c + \mu_n}{c - \mu_n}\right) && (6.7)
\end{aligned}
$$

where (6.6) follows by adding the two positive terms inside the max in (6.5), and (6.7) follows from (6.1) and the fact that $\frac{\beta+\mu_n}{\beta-\mu_n}$ is a decreasing function in β for $\beta > \mu_n$. This completes the proof. □

These two theorems collectively show that it is safe for an individual to contribute data to a database when it is very unlikely to have *any* effect on the public releases of information. This may seem like a simple result, but this is the sort of mechanism that intuitively seems private but which is ruled out by differential privacy.

MAP estimators. The sufficient conditions shown above are simple and they cover some very practical and important functions. As an example, we consider a wide class of estimators known in the literature as "maximum a posteriori probability" (MAP) estimators [62]. At a high level, the scenario we are interested in is one where the database rows are sampled i.i.d. from one of several distributions, but where which distribution is used is not known. The MAP estimator calculates, based on provided prior probabilities of each distribution being used, the distribution from

which the database entries are most likely sampled. The MAP estimator appears a lot in applications involving parameter estimation and multiple hypothesis testing. We show that releasing the MAP estimator achieves the notions of $(0,\ \delta,\ \Delta)$- and $(\epsilon,\ 0,\ \Delta')$-DDP for two (slightly different) large classes of priors Δ and Δ'. As the MAP estimator is deterministic, this is not possible with differential privacy.

Formally, we consider database entries to come from a set $\mathcal{U}$, and we have a finite family of probability density functions $(f_1, \ldots, f_k)$ that each represents a distribution over $\mathcal{U}$. A distribution $\mathcal{D} \in \Delta$ generates the database X as follows. First, $\mathcal{D}$ picks one of the distributions in the family $(f_1, \ldots, f_k)$, with each f_i chosen with probability p_i (for some probability mass function $(p_1, \ldots, p_k)$). Once that choice has been made, the entries of X are chosen i.i.d. according to the chosen f_i. A given $\mathcal{D}$ is defined by the choice of $(p_1, \ldots, p_k)$, and we take Δ to be the union of all distributions $\mathcal{D}$ defined in this manner, where the union is taken over all legitimate probability mass functions $(p_1, \ldots, p_k)$.

A MAP estimator F with respect to $(f_1, \ldots, f_k)$ takes as input a set of prior probabilities $(\pi_1, \ldots, \pi_k)$ (summing to 1) that the user assigns to the distributions $(f_1, \ldots, f_k)$ and outputs the index of the distribution which is most likely to have generated the given database x. Formally, the output is the value of i that maximizes $\pi_i \prod_{j=1}^{n} f_i(x_j)$ (with ties broken arbitrarily). We emphasize that while intuitively the user is trying to match $(\pi_1, \ldots, \pi_k)$ to the actual priors $(p_1, \ldots, p_k)$, we assume no relationship between them when proving privacy. That is, our results cover the case where the actual priors are unknown to the user.

In the following theorem we show that the MAP estimator F, as defined above, is $(0,\ \delta,\ \Delta)$-DDP, where δ decays exponentially to zero in n, if the family $(f_1, \ldots, f_k)$ satisfies some additional regularity conditions.

Theorem 6.3 *Consider a MAP estimator $F : \mathcal{U}^{n+1} \to [k]$ for a given family of distributions $(f_1, \ldots, f_k)$ each with common support and a set of strictly positive user-defined weights $(\pi_1, \ldots, \pi_k)$. Suppose that the distribution family satisfies the following condition:*

$$\exists\ M > 1 \quad s.t. \quad \forall i, i' \in [k], \forall a \in \mathsf{Supp}(f_i),$$
$$f_i(a) \leq M f_{i'}(a) \tag{6.8}$$

Then the MAP estimator F is $(0,\ \delta,\ \Delta)$-DDP where Δ is defined as above (with the same choice of distribution family $(f_1, \ldots, f_k)$) and

$$\delta = (k-1)\,\tau\,(M+1)\,e^{-nu} \tag{6.9}$$

where

$$\tau = \max_{i \neq i'} \frac{\pi_i}{\pi_{i'}} \tag{6.10}$$
$$u = \min_{i \neq i'} \left(-\ln \left(E\left[\left(\frac{f_i(X_1)}{f_{i'}(X_1)} \right)^s \right] \right) \right) \tag{6.11}$$

for any fixed $s \in (0, 1)$.

To prove this theorem, we first give the following two lemmas.

Lemma 6.1 *Suppose that the family $(f_1, \ldots, f_k)$ satisfies condition 6.8 above and that it contains no two members which are essentially the same. That is, we have*

$$\Pr\left[f_i(Y) \neq f_{i'}(Y) \mid i\right] > 0 \qquad \forall i \neq i' \tag{6.12}$$

where $\Pr[\cdot \mid i]$ is the probability computed with Y drawn according to f_i. (Below we use similar notation with expected value.) Let $s \in (0,1)$. Then, we must have

$$0 < \quad E\left[\left(\frac{f_{i'}(Y)}{f_i(Y)}\right)^s \mid i\right] < 1 \qquad \forall i \neq i'$$

Proof: Suppose that $E\left[\left(\frac{f_{i'}(Y)}{f_i(Y)}\right)^s \mid i\right] = 0$. This requires that $f_{i'}(a) = 0$ for all a that occur with positive probability under f_i, but this is impossible because it would violate condition 6.8. Next, by Jensen's inequality, we have

$$E\left[\left(\frac{f_{i'}(Y)}{f_i(Y)}\right)^s \mid i\right] \quad < \left(E\left[\frac{f_{i'}(Y)}{f_i(Y)} \mid i\right]\right)^s \leq 1$$

where the first inequality is strict due to the fact that condition (6.12) holds and the fact that $s \in (0,1)$. □

Lemma 6.2 *Fix some $j \in [k]$. Let Δ contain the single distribution where the entries of the database X are drawn i.i.d. according to f_j. Then, the MAP estimator F defined above is $(0,\ \delta,\ \Delta)$-DDP where δ is as given by (6.9).*

Proof: Let $i \in [n]$, $x_i \in \mathsf{Supp}(f_j)$, and $s \in (0,1)$. For all $j' \neq j$, define

$$u_{j',j}(s) = -\ln\left(E\left[\left(\frac{f_{j'}(X_1)}{f_j(X_1)}\right)^s\right]\right)$$

where the expectation is with respect to X_1 drawn according to f_j. Note that, using Lemma 6.1, it follows that $u_{j',j}(s) > 0 \; \forall \; s \in (0,1), \; \forall \; j' \neq j$. We give upper bounds on $\Pr\left[F(X) \neq j \mid X_i = x_i\right]$ and $\Pr\left[F(X_{-i}) \neq j \mid X_i = x_i\right]$ as follows. First, observe that

$$\begin{aligned}
&\Pr\left[F(X) \neq j \mid X_i = x_i\right] \\
&\quad \leq \sum_{j' \neq j} \Pr\left[\pi_{j'} \prod_{\ell \neq i} f_{j'}(X_\ell) f_{j'}(x_i) > \pi_j \prod_{\ell \neq i} f_j(X_\ell) f_j(x_i)\right] \qquad (6.13) \\
&\quad = \sum_{j' \neq j} \Pr\left[\prod_{\ell \neq i} \left(\frac{f_{j'}(X_\ell)}{f_j(X_\ell)}\right)^s > e^{-s\beta_{x_i,j',j}}\right] \\
&\quad \leq \sum_{j' \neq j} e^{s\beta_{x_i,j',j}} e^{-nu_{j',j}(s)} \qquad (6.14) \\
&\quad \leq \left(|\mathcal{K}| - 1\right) \tau M e^{-nu} \qquad (6.15)
\end{aligned}$$

where $\beta_{x_i,j',j} = \ln\left(\frac{\pi_j f_j(x_i)}{\pi_{j'} f_{j'}(x_i)}\right)$ and $\tau, \; M$, and u are as defined in Theorem 6.3 above. Note that (6.13) follows from the union bound and the definition of the MAP estimator F, (6.14) follows from Markov's inequality, and (6.15) follows from the fact that $e^{s\beta_{x_i,j',j}} \leq \tau M$ and $u \leq u_{j',j}(s)$ for all $j' \neq j, \; , x_i \in \mathsf{Supp}(f_j), \; s \in (0,1)$.

Similarly,

$$\begin{aligned}
\Pr\left[F(X_{-i}) \neq j \mid X_i = x_i\right] &\leq \sum_{j' \neq j} \Pr\left[\pi_{j'} \prod_{\ell \neq i} f_{jr'}(X_\ell) > \pi_j \prod_{\ell \neq i} f_j(X_\ell)\right] \\
&= \sum_{j' \neq j} \Pr\left[\prod_{\ell \neq i} \left(\frac{f_{j'}(X_\ell)}{f_j(X_\ell)}\right)^s > e^{-s\tilde{\beta}_{j',j}}\right] \\
&\leq \sum_{j' \neq j} e^{s\tilde{\beta}_{j',j}} e^{-nu_{j',j}(s)} \\
&\leq (|\mathcal{K}| - 1)\, \tau e^{-nu} \qquad (6.16)
\end{aligned}$$

where $\tilde{\beta}_{j',j} = \ln\left(\frac{\pi_j}{\pi_{j'}}\right)$, and (6.16) follows from the fact $e^{s\tilde{\beta}_{j',j}} \leq \tau$ and $u \leq u_{j',j}(s)$ for all $j' \neq j,\ s \in (0,1)$.

Now, observe that

$$\begin{aligned}
&\Pr\left[F(X) = F(X_{-i}) \mid X_i = x_i\right] \geq \Pr\left[F(X) = F(X_{-i}) = j \mid X_i = x_i\right] \\
\Rightarrow &\Pr\left[F(X) \neq F(X_{-i}) \mid X_i = x_i\right] \leq \Pr\left[F(X) \neq j \mid X_i = x_i\right] \\
&\qquad + \Pr\left[F(X_{-i}) \neq j \mid X_i = x_i\right] \qquad (6.17)
\end{aligned}$$

The result follows directly from (6.15), (6.16), and (6.17) together with Theorem 6.1.

□

The proof of Theorem 6.3 is straightforward at this point. Theorem 5.4 of Section 5.4, together with Lemma 6.2 above, extends the distributional differential privacy of the MAP estimator F from any distribution f_j in the family $(f_1, \ldots, f_k)$ to the convex hull of $\{f_1, \ldots, f_k\}$ which is indeed the required Δ.

Next, we consider another class of priors Δ_λ indexed by some $\lambda > 0$. This class is defined equivalently to Δ, with the added condition that $p_i \geq \lambda$ for all i. For this class of priors, we give the following result that asserts that the MAP estimator F, as defined above, is $(\epsilon,\ 0,\ \Delta_\lambda)$-DDP, where ϵ decays to zero in n, if the family $(f_1, \ldots, f_k)$ satisfies the same regularity conditions as in the previous theorem.

Theorem 6.4 *Consider a MAP estimator $F : \mathcal{U}^{n+1} \to [k]$ for a given family of distributions $(f_1, \ldots, f_k)$ each with common support and a set of strictly positive user-defined weights $(\pi_1, \ldots, \pi_k)$. Suppose that the distribution family satisfies the following condition:*

$$\exists\ M > 1 \quad s.t. \quad \forall i, i' \in [k], \forall a \in Supp(f_i),$$
$$f_i(a) \leq M f_{i'}(a) \tag{6.18}$$

Then the MAP estimator F is $(\epsilon,\ 0,\ \Delta_\lambda)$-DDP where

$$\epsilon = \ln\left(\frac{1 + (M/\lambda - 1)\,\tau M e^{-nu}}{1 - (k-1)\,\tau M e^{-nu}}\right) \tag{6.19}$$

where τ is given by (6.10) and u is given by (6.11) for any fixed $s \in (0, 1)$.

Proof: Let $\mathcal{D} \in \Delta_\lambda$ be the distribution of the database X. Hence, $\mathcal{D}$ can be written as a convex mixture of the members of $\{f_1, \ldots, f_k\}$ where the coefficients of such mixture is given by some probability mass function $p_j,\ j \in [k]$ that satisfies $p_j \geq \lambda\ \forall j \in [k]$. Let $i \in [n]$, $j \in [k]$, and $x_i \in \mathrm{Supp}(X_i)$. Note that, due to condition

(6.8), x_i lies in the common support of $\{f_j,\ j \in [k]\}$. Define

$$\begin{aligned} \mathrm{Num}(i,j,x_i) &= \Pr[F(X) = j \mid X_i = x_i] \\ \mathrm{Den}(i,j,x_i) &= \Pr[F\left(X_{-i}\right) = j \mid X_i = x_i] \end{aligned}$$

We need to show that $e^{-\epsilon} \leq \frac{\mathrm{Num}(i,j,x_i)}{\mathrm{Den}(i,j,x_i)} \leq e^{\epsilon}$ where ϵ is given by (6.19). We will use the notation $\Pr[\cdot \mid \bar{j}]$ to denote the probability of an event conditioned on the fact that f_j is *not* selected as the database-generating distribution. First, observe that we can write

$$\begin{aligned} \mathrm{Num}(i,j,x_i) &= \Pr[F(X) = j \mid j,\ X_i = x_i]\hat{p}_{j|x_i} \\ &\quad + \Pr[F(X) = j \mid \bar{j}, X_i = x_i](1 - \hat{p}_{j|x_i}) \\ \mathrm{Den}(i,j,x_i) &= \Pr[F\left(X_{-i}\right) = j \mid j]\hat{p}_{j|x_i} \\ &\quad + \Pr[F\left(X_{-i}\right) = j \mid \bar{j}, X_i = x_i](1 - \hat{p}_{j|x_i}) \end{aligned} \tag{6.20}$$

where

$$\hat{p}_{j|x_i} = \frac{p_j f_j(x_i)}{\sum_{j' \in [k]} p_{j'} f_{j'}(x_i)}$$

Note that, using the lower bound on p_j and condition (6.8), $\hat{p}_{j|x_i}$ can be lower bounded as

$$\hat{p}_{j|x_i} \geq \frac{\lambda}{M} \tag{6.21}$$

Note also that we dropped the conditioning on $X_i = x_i$ in the first term on the right-hand side of (6.20) since, conditioned on f_j being the database-generating distribution, X_{-i} and X_i are independent.

Let $s \in (0,1)$. Define

$$u_{j',j}(s) = -\ln\left(E\left[\left(\frac{f_{j'}(X_1)}{f_j(X_1)}\right)^s \mid j\right]\right)$$

Using Lemma 6.1, it follows that $u_{j',j}(s) > 0 \ \forall \ s \in (0,1), \ \forall \ j, j' \in [k]$ such that $j \neq j'$. Following similar steps to those that lead to the derivation of (6.15), we have

$$\begin{aligned} 1 \geq \Pr[F(X) = j \mid j, X_i = x_i] &\geq 1 - \sum_{j' \neq j} e^{s\beta_{x_i,j',j}} e^{-nu_{j,j'}(s)} \\ &\geq 1 - (\mathcal{K} - 1)\,\tau M e^{-nu} \end{aligned} \tag{6.22}$$

where $\beta_{x_i,j',j} = \ln\left(\frac{\pi_j f_j(x_i)}{\pi_{j'} f_{j'}(x_i)}\right)$. Using similar analysis, we get

$$1 \geq \Pr\left[F\left(X_{-i}\right) = j \mid j, X_i = x_i\right] \geq 1 - (\mathcal{K} - 1)\,\tau e^{-nu} \tag{6.23}$$

On the other hand, we have $0 \leq \Pr\left[F(X) = j \mid \bar{j}, X_i = x_i\right]$ and

$$\begin{aligned}
&\Pr\left[F(X) = j \mid \bar{j}, X_i = x_i\right] \\
&\quad = \Pr\left[\pi_j \prod_{\ell=1}^{n} f_j(X_\ell) \geq \max_{\hat{j}:\hat{j}\neq j} \pi_{\hat{j}} \prod_{\ell=1}^{n} f_{\hat{j}}(X_\ell) \mid \bar{j}, X_i = x_i\right] \\
&\quad = \sum_{j'\neq j} \Pr\left[\pi_j \prod_{\ell=1}^{n} f_j(X_\ell) \geq \max_{\hat{j}:\hat{j}\neq j} \pi_{\hat{j}} \prod_{\ell=1}^{n} f_{\hat{j}}(X_\ell) \mid j', X_i = x_i\right] \frac{\hat{p}_{j'|x_i}}{1-\hat{p}_{j|x_i}} \\
&\quad \leq \sum_{j'\neq j} \Pr\left[\pi_j \prod_{\ell=1}^{n} f_j(X_\ell) \geq \pi_{j'} \prod_{\ell=1}^{n} f_{j'}(X_\ell) \mid j',\ X_i = x_i\right] \frac{\hat{p}_{j'|x_i}}{1-\hat{p}_{j|x_i}} \\
&\quad \leq \tau M e^{-nu} \qquad (6.24)
\end{aligned}$$

where (6.24) follows from Markov's inequality, (6.8), (6.10), and (6.11). Using similar analysis, we get

$$0 \leq \Pr\left[F\left(X_{-i}\right) = j \mid \bar{j}, X_i = x_i\right] \leq \tau e^{-nu} \qquad (6.25)$$

From (6.21), (6.22), (6.23), (6.24), and (6.25), we have

$$\left(1-(\mathcal{K}-1)\,\tau M e^{-nu}\right)\frac{\lambda}{M} \quad \leq \mathrm{Num}(i,j,x_i) \leq 1+\tau M e^{-nu}\left(1-\tfrac{\lambda}{M}\right)$$

and

$$\left(1-(\mathcal{K}-1)\,\tau e^{-nu}\right)\frac{\lambda}{M} \leq \mathrm{Den}(i,j,x_i) \leq 1+\tau e^{-nu}\left(1-\frac{\lambda}{M}\right)$$

Hence, it is easy to see that $e^{-\epsilon} \leq \frac{\text{Num}(i,j,x_i)}{\text{Den}(i,j,x_i)} \leq e^{\epsilon}$ for all $i \in [n],\ j \in [k]$, and $x_i \in \text{Supp}(X_i)$. □

The above results hold for a distribution class with no auxiliary information, but they can be extended to the case where the auxiliary information Z is given by a proper subset of the database entries $X_{\mathcal{L}} = \{X_j,\ j \in \mathcal{L} \subset [n]\}$. Theorem 5.5 lets us treat this as an upper bound, meaning the result applies for any $Z = g(X_{\mathcal{L}})$ (where g is a randomized function). The proof follows exactly the same lines of the proofs of Theorems 6.3 and 6.4 after removing the compromised entries from the database.

Corollary 6.1 *Let the auxiliary information be given by $Z = g(X_{\mathcal{L}})$ for any subset $\mathcal{L} \subset [n]$ and any (randomized) function g. The results of Theorems 6.3 and 6.4 still hold with n in (6.9) and (6.19) being replaced with $n - L$ where $L = |\mathcal{L}|$.*

We believe that both the sufficient conditions and the MAP estimator mechanism itself are of interest independent of our privacy definition. Because the databases' inherent randomness here is only used to avoid problematic situations, rather than as a substitute for added noise, we believe a similar (though possibly slightly less utile) mechanism could be shown to be private under other privacy definitions as well. Mainly, one can show that with a little added noise, the MAP mechanism can be made ϵ-differentially private.

6.2 Histograms

Sampling distributions, in which the data are drawn randomly from a fixed underlying population, form a natural class of distributions on data sets. We argue that a *truncated histogram*, which releases a histogram (or contingency table) from which small cell counts have been redacted, is DDP for a large subclass of sampling distributions (and their convex combinations).

The model here is that the random sample is the input to the mechanism. In this context, distributional differential privacy ensures that an adversary cannot learn about an individual, *even if the attacker knows that the individual was in the sample*. Consequently, the adversary cannot determine if a given individual was in the sample to begin with. Our results strengthen results of Gehrke et al. [29] on truncated histograms; we explain the relationships between the results further below.

The only condition we require on the sampling distribution is that the size of the sample (denoted N because it is now a random variable) has some uncertainty (to the adversary).

Definition 6.1 (Sampling Priors) *Given a finite multiset P (the "population"), and a distribution p_N on nonnegative integers, the* sampling distribution $\mathcal{D}_{P,p_N}$ *picks N according to p_N and obtains X by selecting (without replacement) N individuals uniformly at random from the population P.*

The class $\Delta_{(\epsilon,\delta)\text{-}Samp}$ is the convex closure of the set of sampling distributions

for which the random variable N satisfies

$$\Pr_{n\sim p_N}\left(e^{-\epsilon} \leq \frac{\Pr(N=n)}{\Pr(N=n-1)} \leq e^{\epsilon}\right) \geq 1-\delta\,. \tag{6.26}$$

The condition on the randomness of the sample size holds in a variety of settings. It is slightly stronger than requiring $N \approx_{\epsilon,\delta} N+1$—it corresponds to requiring that N and $N+1$ be "pointwise" (ϵ,δ)-indistinguishable in the terminology of [45]. Nevertheless, N satisfies the condition when N is either binomial or Poisson[1] (as long as the expectation is sufficiently large, see below) or when $N = const + \mathsf{Lap}(1/\epsilon)$ where Lap is the Laplace distribution. The following lemma is useful in both the discussions of priors and the proof of our main result.

Lemma 6.3 *For every $\epsilon, \lambda, p, n > 0$, we have (1) The Poisson distribution $\mathsf{Po}(\lambda)$ satisfies Eq.* (6.26) *when $\delta = \exp(-\lambda\epsilon^2)$, and (2) $Bin(n,p)$ satisfies Eq.* (6.26) *where $\delta = \exp(-\Omega(np\epsilon^2))$.*

Some examples of sampling priors that fall in the class $\Delta_{(\epsilon,\delta)\text{-Samp}}$:

- Suppose the input is obtained by sampling each element in P independently with probability $p \in (0,1)$. The size N of the sample is binomial $Bin(|P|,p)$, and satisfies our condition when $|P|\cdot p$ is $\Omega(\frac{\log(1/\delta}{\epsilon^2})$ (see Lemma 6.3).

- Suppose the input to the mechanism is a sample of some known, fixed size n_0. One can enforce the randomness condition by discarding only a few data points at random: set $N = n_0 - [\mathsf{Lap}(1/\epsilon) + \frac{\log(1/\delta)}{\epsilon}]_+$ points and discard all but

[1] Recall that for any nonnegative real number λ, $\mathsf{Po}(\lambda)$ is the distribution over nonnegative integers such that $P(N=n) = e^{-\lambda}\lambda^n/n!$.

N data points. Here Lap denotes the Laplace distribution and $[x]_+$ denotes $\max\{x, 0\}$. Note that $N \leq n_0$.

This results in a randomized mechanism that alters at most $\frac{2\log(1/\delta)}{\epsilon}$ bin counts in the histogram, far less than the number required to ensure differential privacy (which requires altering the counts of all bins with some probability).

- *Poisson priors:* In Poisson sampling, the sample size N follows a Poisson distribution. It satisfies our condition when λ is $\Omega(\frac{\log(1/\delta)}{\epsilon^2})$.

- The definition is phrased in terms of a fixed population P, but i.i.d. sampling also falls into this class (one obtains i.i.d. sampling in the limit as $|P|$ goes to infinity).

Given a partition of the data domain $\mathcal{U}$ into "bins", a histogram reports the number of data points in each bin. The k-truncated histogram reports the set of counts with value at least k (and reports "0" for counts less than k).

Theorem 6.5 (Privacy via Sampling Priors) *There is a constant $C > 0$ such that, for $k > \frac{C\log(1/\delta)}{\epsilon^2}$, the k-truncated histogram is $(3\epsilon, 3\delta)$-DDP for the class $\Delta_{(\epsilon,\delta)\text{-}Samp}$.*

The main difficulty of the proof is that the histogram counts – that is, the entries of the vector $F(X_i)$ – are not independent. For example, when $N = const + Lap(1/\epsilon)$, then the sum of the counts is much more concentrated than it would be if the entries were truly independent (or even if every single count were close to independent from the remaining ones). Nonetheless, we can use the randomness in

N to limit the information about the jth entry of $F(X_i)$ that is contained in the remaining entries.

Proof of Theorem 6.5: Recall that our goal is to compare $F(X)$ and $F(X_{-i})$, conditioned on element X_i taking a fixed value.

Suppose we condition on X_i lying in bin j. Observe that

$$F(X) = F(X_{-i}) + e_j$$

(where e_j is a vector equal to 1 in position j and 0 elsewhere), so we really need to compare $F(X_i)$ with $F(X_{-i}) + e_j$. For a particular value of $N = n$, there are two cases: First, if the expectation of $F(X)_j$ is very small (less than $k/2$), then the count for bin j will be less than $k-2$ with high probability and so the count for bin j will be suppressed (whether or not e_j is added). We omit the calculation.

In the second case, the expected count for bin j is large. Let M denote this count, which is binomially distributed for a particular value of N. Since $E(M)$ is large, we have $M \approx_{\epsilon,\delta} M + 1$ (by Lemma 6.3). Unfortunately, this is not quite enough, since the bin counts are not independent (nor even close to independent), and we cannot analyze M on its own.

We'll say a value n of N is "good" if $\frac{\Pr(N=n)}{\Pr(N=n-1)} \in e^{\pm\epsilon}$. *For now, we condition on a particular good value n.* Because X consists of a uniform sample of size N, we have that the distributions of $F(X_{-i})\big|_{N=n}$ and $F(X)\big|_{N=n-1}$ are identical. Since n is "good", it suffices to compare the distributions of $F(X)$ with $F(X) + e_j$. The entries of F are not independent in general, so we need to analyze the vectors as a

whole.

Our strategy is to "couple" these two random variables to make the comparison easier. We therefore compare $F(X)|_{N=n}$ with $(F(X) + e_j)|_{N=n-1}$. Because n is "good", we can later account for the discrepancy between the two events we condition on.

Next, we wish to isolate the information in the j-th entry. Let M denote the j-th entry of $F(X)$ (the count of bin j). Conditioned on the value of $N - M$ (i.e., the sum of the remaining counts), then the vector of remaining counts is independent of M, so it suffices to consider the distribution on the pair $(M, N - M)$. Comparing $F(X)|_{N=n}$ with $(F(X) + e_j)|_{N=n-1}$ amounts to comparing the distributions on $(M, n - M)|_{N=n}$ and $(M + 1, (n - 1) - M)|_{N=n-1}$.

The probability under these two distributions of a pair $(m, n-m)$ can now be computed explicitly. Suppose that $p|P|$ elements of P lie in bin j. Then

$$\Pr(M = m \mid N = n) = \frac{\binom{p|P|}{m}\binom{(1-p)|P|}{n-m}}{\binom{|P|}{n}} \quad \text{and} \quad \Pr(M{+}1 = m \mid N = n{-}1) = \frac{\binom{p|P|}{m-1}\binom{(1-p)|P|}{n-m}}{\binom{|P|}{n-1}}$$

The ratio of these two probabilities is thus $\frac{n}{m} \cdot p \cdot \frac{|P|-m/p}{|P|-n}$. This ratio is $e^{\pm 2\epsilon}$ when $\frac{m}{n}$ is $e^{\pm\epsilon}$. By a multiplicative Chenroff bound, this latter event happens with probability at least $1 - \delta$ because $E(M) \geq k/2$ (where $k = \Omega(\log(1/\delta)/\epsilon^2)$). This completes the analysis for "good" values of n. By assumption, N is good with probability at least $1 - \delta$. □

We now prove the remaining necessary lemmas.

Proof of Lemma 6.3: For Part 1, fix an integer $m > 0$. If $N \sim \mathsf{Po}(\lambda)$, the ratio

$\frac{\Pr(N=m)}{\Pr(N=m-1)} = \frac{\lambda}{m}$. Thus, to compare $\mathsf{Po}(\lambda)$ and $1 + \mathsf{Po}(\lambda)$, it suffices to compute the probability that N lies outside of $[\lambda e^{-\epsilon}, \lambda e^{\epsilon}]$. By standard tail bounds (Lemma 6.4), this probability is $\exp(-\Omega(\lambda\epsilon^2))$. Part 2 follows a similar calculation, replacing the Poisson tail bound with the multiplicative Chernoff bound. □

We use the following tail bound, which follows from the definition of the Poisson distribution and Stirling's approximation for the factorial.

Lemma 6.4 (Poisson Tail Bounds) *If* $N \sim \mathsf{Po}(\lambda)$, *then* $\max\{\Pr(N > \lambda e^{\epsilon}), \Pr(N < \lambda e^{-\epsilon})\} \leq \frac{1}{\epsilon\sqrt{2\pi}} \exp(-\frac{\lambda\epsilon^2}{2}(1 + o(\epsilon)))$.

Relation to the work of Gehrke et al. Gehrke et al. [29] prove a result which appears, at first glance, very similar: namely, that a mechanism which samples each input record with probability p and computes a histogram on the resulting sample is differentially private.

There are two principal differences between the results. First, we assume the sample *is* the database, and so we ask that the adversary not be able to learn about a particular individual *compared to a world where they were not in the sample*. This corresponds very closely, for example, to preventing the type of attack carried out by Homer et al. [37] on genome-wide association study data. In contrast, Gehrke et al. treat the population as the database. This means that the adversary will not be able to learn about a particular individual *compared to a world where they were not in the population*, where some of the privacy comes from the fact that the individual was probably not sampled in the first place.

Second, the parameters of the two results are incomparable: Gehrke et al. as-

sume the sample itself is very small—approximately an ϵ fraction of the population—whereas our results apply to populations that are very close in size to N (subject to the population always being larger than N). On the other hand, Gehrke et al. require only that k be approximately $\log(1/\delta)/\epsilon$, instead of $\log(1/\delta)/\epsilon^2$. The bound on k is tight for our definition, unfortunately.

Finally, we mention that Gehrke et al. introduce a new definition, called *crowd-blending privacy*. This definition requires that the data of an individual could be replaced with the data of some number of other individuals in the database without much change in the output distribution. There is no inference-based version of the definition given, and whether there are situations where crowd-blending privacy gives sufficient privacy guarantees is left unclear. The primary motivation, however, is that when a crowd-blending private query is run on a database sampled from a larger population, the sample-then-query combination as a whole is differentially private when considering the population as a whole to be the database. (This is how the result mentioned above is achieved.)

It seems likely to us that the result discussed above for histograms could be generalized to give a way of converting crowd-blending private mechanisms to DDP mechanisms, though we have not verified this. Doing so, or comparing to other instantiations of coupled-worlds privacy, could add to the understanding of what exactly crowd-blending privacy protects about an individual and in what situations.

6.3 Sums

In this section, we consider a mechanism that releases the sum (or equivalently, average) of the entries in a real-valued database without any form of randomization or added noise. We begin with a simple warm-up case, a distribution consisting of rows each chosen i.i.d. from a uniform distribution on $[0, 1]$ with no auxiliary information. We show that a given row X_i is effectively hidden by the noise of the other added rows, the sum of which is distributed according to the Irwin-Hall distribution, which is a close approximation of a Gaussian.

Having shown this initial result, we then generalize substantially. We first allow for more general distributions on rows. In particular we consider distributions that contain a *rectangle*, by which we mean the probability density function p is lower-bounded by h on some interval $[s, t]$. (This is generalized to include databases with rows in $\mathbb{R}^d$ rather than just $\mathbb{R}$.) The size and shape of the rectangle has a large impact on the values of ϵ and δ, with the best case occurring when the rectangle has volume close to 1 and has a width close to the full support of the underlying distribution. This generalization requires repeating the proof, with the new proof becoming substantially more complex (but using the same underlying methods).

We then use convenient properties of distributional differential privacy to generalize further. We consider a case where auxiliary information discloses some of the rows of the database. We show that the previous theorem holds, with the number of undisclosed rows taking the place of n. Theorem 5.5 then shows that privacy is maintained when instead of full disclosure, the auxiliary information is simply some

function of the previously disclosed rows. This means that what we really assume is that the adversary's auxiliary information is independent of some number of rows, an upper bound that is certainly reasonable in many situations.

We then use Theorem 5.4 to remove the assumption that rows are independent of each other. We assume instead that the rows are drawn i.i.d. from *some* distribution, but the adversary's prior (or the real world) might put positive probability on each of many possible distributions. This means that it is acceptable to release an average even when that average really does constitute new information for the adversary or the public.

This result is being presented in place of an earlier result that also showed sums could be released privately [3]. This result achieved better parameters and applies to a wider class of distributions. The application to linear regression in the following section would also be difficult at best using the earlier result.

6.3.1 Background

Here we introduce some background and notation. In particular, we need to use the Irwin-Hall distribution, given below.

Definition 6.2 *The* Irwin-Hall distribution *is the distribution obtained by adding n i.i.d. random variables, each drawn from the uniform distribution on $[0, 1]$. The probability density function can be written as*

$$pdf_n(y) = \frac{1}{2\,(n-1)!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} (y-k)^{n-1}\, \mathsf{sgn}(y-k)$$

or as

$$pdf_n(y) = \frac{1}{(n-1)!}\sum_{k=0}^{\lfloor y \rfloor}(-1)^k\binom{n}{k}(y-k)^{n-1}.$$

The cumulative density function can be written as

$$cdf_n(y) = \frac{1}{n!}\sum_{k=0}^{\lfloor y \rfloor}(-1)^k\binom{n}{k}(y-k)^{n}.$$

The Irwin-Hall distribution converges very quickly and becomes an extremely good approximation of the Gaussian distribution. As a result, we can show privacy using calculations similar to those used to show that Gaussian noise is an acceptable way to achieve differential privacy.

In our proofs we discuss the quantity $\frac{pdf_n(y)}{pdf_n(y-c)}$ and assume that its largest values occur at the tails of the Irwin-Hall distribution. This is computationally verified and is true for the Gaussian distribution, which is a very close approximation, but the lack of a simple expression for *pdf* makes it difficult to verify formally.

Assumption 6.1 *Consider the quantity* $\frac{pdf_n(y)}{pdf_n(y-c)}$ *for* $0 < y < n/2$ *and* $0 < c < y$. *For a fixed* c, *we assume this quantity is decreasing in* y. *That is, the highest values occur at the tail of the distribution.*

We will also use the following standard technical lemma.

Lemma 6.5 *For any two events* A *and* B,

$$|\Pr[A] - \Pr[A \mid B]| \leq \Pr[\neg B].$$

Proof: First condition:

$$\begin{aligned}\Pr[A] &= \Pr[A \mid B]\Pr[B] + \Pr[A \mid \neg B]\Pr[\neg B] \\ &\leq \Pr[A \mid B] + \Pr[\neg B]\end{aligned}$$

Second condition:

$$\begin{aligned}\Pr[A] &= \Pr[A \mid B]\Pr[B] + \Pr[A \mid \neg B]\Pr[\neg B] \\ &\geq \Pr[A \mid B]\Pr[B] \\ &\geq \Pr[A \mid B](1 - \Pr[\neg B]) \\ &\geq \Pr[A \mid B] - \Pr[A \mid B]\Pr[\neg B]) \\ &\geq \Pr[A \mid B] - \Pr[\neg B])\end{aligned}$$

□

6.3.2 The Simple Case

We begin with a warm-up result, showing privacy when all variables are uniformly distributed.

Theorem 6.6 *Let $\mathcal{D}$ be the generating distribution that outputs a database X of n rows each chosen i.i.d. from the uniform distribution on $[0, 1]$ and $\Delta = \{\mathcal{D}\}$ (with empty auxiliary information). Let F be the mechanism that outputs the exact sum*

of all rows of the database. For all $a \in [0, n]$*,* F *is* $(\epsilon, \delta, \Delta)$*-DDP with*

$$\epsilon = \ln\left(\frac{pdf_{n-1}(a - .5)}{pdf_{n-1}(a - 1)}\right)$$

$$\delta = cdf_{n-1}(a - .5) + cdf_{n-1}(a).$$

Proof: **Step 1:** Without loss of generality, assume that i (from the DDP definition) takes value n. Let $T = F(X)$ be the sum of all rows, and $T' = \mathsf{Sim}(X)$ be the sum of all rows, but with .5 used in place of the unknown x_n. Note that fixing $X_n = c$ the probability density function for T is $pdf_{n-1}(y - c)$ (where pdf_{n-1} refers to the analogous function for the Irwin-Hall distribution with parameter $n-1$ and y is the potential value of T). The probability density function of T' is $pdf_{n-1}(y - .5)$. We want to show that for any set S and any fixed $c \in [0, 1]$,

$$\Pr[T' \in S] \leq e^{\epsilon} \Pr[T \in S] + \delta. \tag{6.27}$$

We pick a constant a in $[0, n]$. This is a parameter that will allow us to make trade-offs between our resulting ϵ and δ values. We divide S into two sets, $\mathsf{Far} = S \cap ((-\infty, a] \cup [n - a, \infty))$ and $\mathsf{Close} = S \cap [a, n - a]$. Far represents the tails, where privacy will fail (and which will be covered by the δ term) and Close represents the bulk of the distribution. We now modify our equation.

$$\Pr[T' \in \mathsf{Far}] + \Pr[T' \in \mathsf{Close}] \leq e^{\epsilon}(\Pr[T \in \mathsf{Far}] + \Pr[T \in \mathsf{Close}]) + \delta.$$

We now decrease the right side and increase the left side and maintain a sufficient condition for privacy.

$$\Pr[T' \in \mathsf{Far}] + \Pr[T' \in \mathsf{Close}] \leq e^{\epsilon} \Pr[T \in \mathsf{Close}] + \delta$$

$$\Pr[T' \leq a \vee T' \geq n - a] + \Pr[T' \in \mathsf{Close}] \leq e^{\epsilon} \Pr[T \in \mathsf{Close}] + \delta$$

We then require that

$$\begin{aligned}
\delta &\geq \Pr[T' \leq a] + \Pr[T' \geq n - a] \\
&\geq cdf_{n-1}(a - .5) + [1 - cdf_{n-1}(n - a - .5)] \\
&\geq cdf_{n-1}(a - .5) + cdf_{n-1}((n - 1) - (n - a - .5)) \\
&\geq cdf_{n-1}(a - .5) + cdf_{n-1}(a - .5) \\
&\geq 2cdf_{n-1}(a - .5).
\end{aligned}$$

With this condition in place, we can reduce our condition to

$$\begin{aligned}
\Pr[T' \in \mathsf{Close}] &\leq e^{\epsilon} \Pr[T \in \mathsf{Close}] \\
\Pr[T' = y] &\leq e^{\epsilon} \Pr[T = y] \text{ for all } y \in \mathsf{Close} \\
\frac{\Pr[T' = y]}{\Pr[T = y]} &\leq e^{\epsilon} \text{ for all } y \in \mathsf{Close} \\
\frac{pdf_{n-1}(y - .5)}{pdf_{n-1}(y - c)} &\leq e^{\epsilon} \text{ for all } y \in \mathsf{Close}
\end{aligned}$$

The quantity $\frac{pdf_{n-1}(y-.5)}{pdf_{n-1}(y-c)}$ is decreasing in y when $c > .5$ and increasing in y when $c < .5$. It also increases as c becomes closer to 0 or 1. This means that the worst

case (for $y \in \mathsf{Close}$) occurs when $y = a$ and $c = 1$ or when $y = n - a$ and $c = 0$. That means we must take ϵ such that

$$\frac{pdf_{n-1}(a - .5)}{pdf_{n-1}(a - 1)} \leq e^{\epsilon} \text{ and } \frac{pdf_{n-1}(n - a - .5)}{pdf_{n-1}(n - a)} \leq e^{\epsilon}$$

These two conditions are equivalent. The points in question are symmetric around the middle ($y = \frac{n-1}{2}$) of the Irwin-Hall distribution with parameter $n - 1$. As a result, equation 6.27, the first privacy condition, is satisfied with

$$\epsilon = \ln\left(\frac{pdf_{n-1}(a - .5)}{pdf_{n-1}(a - 1)}\right), \delta = 2cdf_{n-1}(a - .5).$$

Step 2: We now repeat the same process using the other necessary inequality.

$$\Pr[T \in S] \leq e^{\epsilon} \Pr[T' \in S] + \delta. \tag{6.28}$$

We divide S as before into Far and Close, giving us the new equation

$$\Pr[T \in \mathsf{Far}] + \Pr[T \in \mathsf{Close}] \leq e^{\epsilon}(\Pr[T' \in \mathsf{Far}] + \Pr[T' \in \mathsf{Close}]) + \delta.$$

We now decrease the right side and increase the left side and maintain a sufficient condition for privacy.

$$\Pr[T \in \mathsf{Far}] + \Pr[T \in \mathsf{Close}] \leq e^{\epsilon} \Pr[T' \in \mathsf{Close}] + \delta$$

$$\Pr[T \leq a \vee T \geq n - a] + \Pr[T \in \mathsf{Close}] \leq e^{\epsilon} \Pr[T' \in \mathsf{Close}] + \delta$$

We then require that

$$\begin{aligned}
\delta &\geq \Pr[T \leq a] + \Pr[T \geq n - a] \\
&\geq cdf_{n-1}(a - c) + [1 - cdf_{n-1}(n - a - c)] \\
&\geq cdf_{n-1}(a - c) + cdf_{n-1}((n - 1) - (n - a - c)) \\
&\geq cdf_{n-1}(a - c) + cdf_{n-1}(a - 1 + c) \\
&\geq cdf_{n-1}(a - .5) + cdf_{n-1}(a).
\end{aligned}$$

The last line above comes from the fact that of $a-c$ and $a-1+c$, one will be less than $a - .5$ and the other greater (but less than a). Since $cdf_{n-1}(\cdot)$ is a strictly increasing function, this final condition gives an upper bound. In reality, the function in this small interval will be nearly linear and approximated very well by $2cdf_{n-1}(a - .5)$. With this condition in place, we can reduce our condition to

$$\begin{aligned}
\Pr[T \in \mathsf{Close}] &\leq e^{\epsilon} \Pr[T' \in \mathsf{Close}] \\
\Pr[T = y] &\leq e^{\epsilon} \Pr[T' = y] \text{ for all } y \in \mathsf{Close} \\
\frac{\Pr[T = y]}{\Pr[T' = y]} &\leq e^{\epsilon} \text{ for all } y \in \mathsf{Close} \\
\frac{pdf_{n-1}(y - c)}{pdf_{n-1}(y - .5)} &\leq e^{\epsilon} \text{ for all } y \in \mathsf{Close}
\end{aligned}$$

For roughly the same reasons as those discussed before, the quantity $\frac{pdf_{n-1}(y-c)}{pdf_{n-1}(y-.5)}$ is maximized when $y = a$ and $c = 0$ or when $y = n - a$ and $c = 1$, meaning that we

must take ϵ such that

$$\frac{pdf_{n-1}(a)}{pdf_{n-1}(a-.5)} \leq e^{\epsilon} \text{ and } \frac{pdf_{n-1}(n-a-1)}{pdf_{n-1}(n-a-.5)} \leq e^{\epsilon}$$

These two conditions are equivalent. The points in question are symmetric around the middle ($y = \frac{n-1}{2}$) of the Irwin-Hall distribution with parameter $n-1$. As a result, equation 6.28, the first privacy condition, is satisfied with

$$\epsilon = \ln\left(\frac{pdf_{n-1}(a)}{pdf_{n-1}(a-.5)}\right), \delta = cdf_{n-1}(a-.5) + cdf_{n-1}(a).$$

Having now found ϵ and δ variables separately in order to satisfy equations 6.27 and 6.28, we now take the maximum of two values for each in order to find privacy parameters to satisfy both equations simultaneously. For ϵ, it is Equation 6.27 that requires the higher value, whereas for δ the requirement of Equation 6.28 is greater. That gives the following final required parameters.

$$\epsilon = \ln\left(\frac{pdf_{n-1}(a-.5)}{pdf_{n-1}(a-1)}\right), \delta = cdf_{n-1}(a-.5) + cdf_{n-1}(a).$$

□

6.3.3 Main result

We now move to the more general result. The proof largely proceeds as before. We now allow output in $\mathbb{R}^d$, which complicates calculations but requires no fundamentally new ideas. We also allow non-uniform distributions, which does add

a new layer of complexity to the proof. The way we deal with non-uniform distributions is to use a technique similar to rejection sampling. With some probability, a given row is disclosed to the adversary as part of the auxiliary information. This is done in such a way that *conditioned on having not been disclosed* the row is chosen uniformly (over $[0, 1]$ initially, though this is generalized). The methods of the prior, simpler case can then be applied.

While the goal here is a more general proof, we begin with something still not maximally general, and derive the stronger theorem as a corollary at the end by combining that with several other results.

Theorem 6.7 *Let $\mathcal{D}$ be the generating distribution that outputs a database X of n rows each chosen i.i.d. from a distribution over d-tuples of real numbers, with $X_{i,j}$ representing the j^{th} coordinate of the i^{th} row. Let the probability density function $p(y)$ be such that $p(y) \geq h$ for all $y \in [0, 1]^d$ for some $h > 0$. Let $w_{1,j} = \inf(\mathsf{Supp}(X_{i,j}))$ and $w_{2,j} = \sup(\mathsf{Supp}(X_{i,j}))$ be the lower and upper bounds of j^{th} component of this distribution, with $w_j = w_{2,j} - w_{1,j}$ the width. Let $\Delta = \{\mathcal{D}\}$ (with empty/constant auxiliary information). Let F be the mechanism that outputs the (d-dimensional) exact sum of all rows of the database. F is $(\epsilon, \delta, \Delta)$-DDP with*

$$\epsilon = \sum_j \ln\left(\frac{pdf_{\lceil rhn\rceil-1}(a_j)}{pdf_{\lceil rhn\rceil-1}(a_j - .5w_j)}\right)$$

$$\delta = 2\sum_j (cdf_{\lceil rhn\rceil-1}(a_j + .5w_j)) + (1 + e^{\epsilon})e^{-2nh^2(1-r)^2}$$

for any choices of values $a_j \in [0, n/2]$ and $r \in [0, 1]$.

Proof: **Step 1:** We use the fact that it is sufficient to prove security when Z contains *more* information than is stated in the theorem. In particular, we take Z to be a random variable, correlated with X, computed according to the algorithm below. What Z does is disclose rows outside the $[0,1]^d$ region to the adversary, and discloses rows inside that interval with a probability such that conditioned on not having been disclosed each row has uniform distribution on $[0,1]^d$. Our proof will require that a certain number of rows not be disclosed, and privacy will fail if that condition is not met. (This probability will be included in the δ value.) However, because DDP must be satisfied conditioned on any possible value of Z, we cannot have Z simply disclose too many rows with some small probability. Instead, we require that if the algorithm for Z finds itself in a situation where it would disclose too many rows, it instead computes a new value, unrelated to the current value of X, that has the same distribution Z would have (for random X) if it was not releasing too many rows. This means that, for any value of Z, the probability that this value was output as a result of the Z algorithm failing is equal. The algorithm for computing Z is given formally below:

Most of our proof will assume $\lceil rhn \rceil$ rows remain random to the adversary, and that there was no failure. A higher value of r allows more random rows, resulting in a better ϵ. On the other hand, a higher r value results in a great probability that this assumption will fail, resulting in a greater δ.

The likelihood of any particular value of Z is equal regardless of whether a failure case occurred. As a result, when showing that the privacy condition is

Auxiliary Information

For a given X that was drawn from distribution $\mathcal{D}$ and for constants h, d and r, the associated Z will have n rows $Z_1, \ldots, Z_n$, chosen according to the following process:

Step 1:

- If $X_i \notin [0,1]^d$, $Z_i = X_i$.
- If $X_i \in [0,1]^d$, then with probability $h/p(x_i)$ choose $Z_i = \perp$, otherwise $Z_i = X_i$.
- If the number of rows equal to $\perp$ is at least rhn, change a random subset to the respective X_i values so that there are exactly $\lceil rhn \rceil$ rows with value $\perp$ and then terminate successfully. Otherwise, we consider this to be a "failure case" and continue to step 2.

Step 2:

- Draw a new database X from $\mathcal{D}$. Repeat the process from Step 1 using this new database.
- If the number of rows equal to $\perp$ is again less than rhn, repeat the above until it is not.

Figure 6.1: Formal specification of the variable Z.

satisfied (which must be done for an arbitrary, fixed $Z = z$) we can use a single, consistent $\Pr[\mathsf{failure}]$, without the need to condition on $Z = z$.

Note also that, conditioned on $Z = z$, the coordinates of a particular (undisclosed) row are independent of each other. This means that the values of the output $F(X)$ in each coordinate, once the sum of the disclosed rows has been subtracted, will be independent of each other.

Our proof is similar to the earlier proof, and we begin the same way. Without loss of generality, we will again assume that i (from the DDP definition) takes value n. We will also assume this is a $\perp$ row, but also that it could take any value in $\mathsf{Supp}(X_n)$. By doing this, we are treating the problem as if two mutually exclusive worst cases are occurring simultaneously. If x_i was not a $\perp$ row, we would have one

more row of randomness, improving privacy parameters slightly. If x_i was limited to $[0, 1]$ its range, and hence the amount of noise needed to hide its value, would be reduced. In reality, one of these two improvements in parameters must always be possible, but for simplicity we ignore this.

The number of undisclosed rows, $\lceil rhn \rceil$, will function as the number of total rows n did in the earlier proof. For convenience, we use $m = \lceil rhn \rceil$.

Step 2: We must show that the privacy condition is satisfied under any fixed value of $Z = z$ and any fixed row $X_n = c$. $F(X)$ is the simple sum of all rows (when thought of as vectors in d-dimensional space). For $\mathsf{Sim}(X)$ we proceed similarly to the previous proof, using a function that takes the sum as usual but replaces the missing row with a value $\tilde{w}$ representing the "middle" possible value. Formally, the j^{th} coordinate $\tilde{w}_j$ of $\tilde{w}$ has value $(w_{2,j} - w_{1,j})/2$.

We want to show that for any set S, any fixed $c \in \mathsf{Supp}(X_n)$, and any choice of z,

$$\Pr[\mathsf{Sim}(X) \in S] \leq e^{\epsilon} \Pr[F(X) \in S] + \delta. \tag{6.29}$$

All probabilities are conditioned on $Z = z$, so we drop this notation for convenience. We cannot make guarantees about the situation where a failure occurred during the generation of z, so we use Lemma 6.5 to switch the probabilities above so that they are now conditioned on $\mathsf{success}$. In doing so, these probabilities are altered by at most $\Pr[\mathsf{fail}]$. We assume the worst case about this change so that we still have a

sufficient condition.

$$\Pr[\mathsf{Sim}(X) \in S \mid \mathsf{success}] + \Pr[\mathsf{fail}] \leq e^{\epsilon}(Pr[F(X) \in S \mid \mathsf{success}] - \Pr[\mathsf{fail}]) + \delta$$

$$\Pr[\mathsf{Sim}(X) \in S \mid \mathsf{success}] \leq e^{\epsilon} Pr[F(X) \in S \mid \mathsf{success}] - (1 + e^{\epsilon}) \Pr[\mathsf{fail}] + \delta$$

We take $\delta' = \delta - (1+e^{\epsilon}) \Pr[\mathsf{fail}]$, and we assume all probabilities shown are implicitly conditioned on $\mathsf{success}$. We then have the following condition.

$$\Pr[\mathsf{Sim}(X) \in S] \leq e^{\epsilon} \Pr[F(X) \in S] + \delta'$$

Furthermore, conditioned on $\mathsf{success}$ and on $Z = z$, the value of the sum of all disclosed rows is now fixed. We call this value t. We use T and T' as random values defined analogously to how they were defined in the previous proof. Here T represents the sum of *only* the rows that have value $\perp$ in z. T' represents the sum of these rows, but with $\tilde{w}$ used in place of x_n. This means that $F(X) = T + t$ and $\mathsf{Sim}(X) = T' + t$. Our sufficient condition can now be written as

$$\Pr[T' + t \in S] \leq e^{\epsilon} \Pr[T + t \in S] + \delta'.$$

Since t is fixed, we can use S' to represent the set where all elements of S have been reduced by t and now must show that

$$\Pr[T' \in S'] \leq e^{\epsilon} \Pr[T \in S'] + \delta'.$$

As before, we will divide S' into two sets, Far and Close, with Far representing points far from the middle ($\tilde{w}$) of the distribution, and Close representing the remaining points near the middle. Previously, we used a parameter a to represent the divide between these sets. This time we allow this value to be different in each dimension, with a_j being used to differentiate near from far in the j^{th} dimension. We first define a box $B = \prod_j [a_j + \tilde{w}_j, m - 1 - a_j + \tilde{w}_j]$ and then set $\mathsf{Close} = B \cap S'$. We define Far to be the remaining points, $\mathsf{Far} = S' - \mathsf{Close}$. We now modify our equation.

$$\Pr[T' \in \mathsf{Far}] + \Pr[T' \in \mathsf{Close}] \leq e^{\epsilon}(\Pr[T \in \mathsf{Far}] + \Pr[T \in \mathsf{Close}]) + \delta'.$$

We now decrease the right side and increase the left side and maintain a sufficient condition for privacy.

$$\Pr[T' \in \mathsf{Far}] + \Pr[T' \in \mathsf{Close}] \leq e^{\epsilon} \Pr[T \in \mathsf{Close}] + \delta'$$

We want to require that $\delta > \Pr[T' \in \mathsf{Far}]$. In order to quantify this, we evaluate $\Pr[T' \in \mathsf{Far}]$ by taking a union bound, adding up the probabilities that each

coordinate of T' is outside the interval $[a_j + \tilde{w}_j, m - 1 - a_j + \tilde{w}_j]$.

$$\begin{aligned}
\delta' \geq \Pr[T' \in \mathsf{Far}] &\geq \sum_j \Pr[T'_j \notin [a_j + \tilde{w}_j, m - 1 - a_j + \tilde{w}_j]] \\
&\geq \sum_j (\Pr[T'_j \leq a_j + \tilde{w}_j] + \Pr[T'_j \geq m - 1 - a_j + \tilde{w}_j]) \\
&\geq \sum_j (cdf_{m-1}(a_j) + [1 - cdf_{m-1}(m - 1 - a_j)]) \\
&\geq \sum_j (cdf_{m-1}(a_j) + cdf_{m-1}((m-1) - (m - 1 - a_j))) \\
&\geq \sum_j (cdf_{m-1}(a_j) + cdf_{m-1}(a_j)) \\
&\geq 2 \sum_j cdf_{m-1}(a_j)
\end{aligned}$$

With this condition on δ' in place, we can reduce our condition to

$$\begin{aligned}
\Pr[T' \in \mathsf{Close}] &\leq e^\epsilon \Pr[T \in \mathsf{Close}] \\
\Pr[T' = y] &\leq e^\epsilon \Pr[T = y] \text{ for all } y \in \mathsf{Close} \\
\frac{\Pr[T' = y]}{\Pr[T = y]} &\leq e^\epsilon \text{ for all } y \in \mathsf{Close} \\
\frac{\prod_j \Pr[T'_j = y_j]}{\prod_j \Pr[T_j = y_j]} &\leq e^\epsilon \text{ for all } y \in \mathsf{Close} \\
\prod_j \frac{\Pr[T'_j = y_j]}{\Pr[T_j = y_j]} &\leq e^\epsilon \text{ for all } y \in \mathsf{Close} \\
\prod_j \frac{pdf_{m-1}(y_j - \tilde{w}_j)}{pdf_{m-1}(y_j - c_j)} &\leq e^\epsilon \text{ for all } y \in \mathsf{Close}
\end{aligned}$$

The quantity $\frac{pdf_{m-1}(y_j - \tilde{w}_j)}{pdf_{m-1}(y_j - c_j)}$ is decreasing in y_j when $c_j > \tilde{w}_j$ and increasing in y_j when

$c_j < \tilde{w}_j$. The quantity also increases in each case as c_j becomes more extreme (i.e., as c_j becomes higher when $c_j > \tilde{w}_j$, as c_j becomes lower otherwise). This means that the worst case (for $y \in \mathsf{Close}$) occurs when $y_j = a_j + \tilde{w}_j$ and $c_j = w_{2,j}$ or when $y_j = m - 1 - a_j + \tilde{w}_j$ and $c_j = w_{1,j}$. As seen before, these two cases are symmetric - they simply represent the equivalent situation at the two tales of the Irwin-Hall distribution. Picking one arbitrarily, we must take ϵ such that

$$\prod_j \frac{pdf_{m-1}(a_j + \tilde{w}_j - \tilde{w}_j)}{pdf_{m-1}(a_j + \tilde{w}_j - w_{2,j})} \leq e^{\epsilon}$$

$$\prod_j \frac{pdf_{m-1}(a_j)}{pdf_{m-1}(a_j - .5w_j)} \leq e^{\epsilon}$$

We now know that the first privacy condition, Equation 6.29, is satisfied with

$$\epsilon = \sum_j \ln\left(\frac{pdf_{m-1}(a_j)}{pdf_{m-1}(a_j - .5w_j)}\right)$$

$$\delta' = 2\sum_j cdf_{m-1}(a_j).$$

Step 3: We must now show that we also satisfy the other inequality needed for privacy. Specifically, we want to show that for any set S, any fixed $c \in \mathsf{Supp}(X_n)$, and any choice of z,

$$\Pr[F(X) \in S] \leq e^{\epsilon} \Pr[\mathsf{Sim}(X) \in S] + \delta. \tag{6.30}$$

Again, we want to limit ourselves to $\mathsf{success}$ cases in the generation of z. We can do this exactly the sme way as before. Keeping $\delta' = \delta - (1 + e^{\epsilon}) \Pr[\mathsf{fail}]$, we now have the

following (again, we now assume all probabilities shown are implicitly conditioned on $\mathsf{success}$).

$$\Pr[F(X) \in S] \leq e^{\epsilon} \Pr[\mathsf{Sim}(X) \in S] + \delta'$$

Furthermore, conditioned on z and on $\mathsf{success}$, we again can shift S by t, the sum of all disclosed rows, so that the sufficient condition is now

$$\Pr[T \in S'] \leq e^{\epsilon} \Pr[T' \in S'] + \delta'.$$

Taking the same definitions of Far and Close, this condition becomes

$$\Pr[T \in \mathsf{Far}] + \Pr[T \in \mathsf{Close}] \leq e^{\epsilon}(\Pr[T' \in \mathsf{Far}] + \Pr[S' \in \mathsf{Close}]) + \delta'.$$

We now decrease the right side and increase the left side and maintain a sufficient condition for privacy.

$$\Pr[T \in \mathsf{Far}] + \Pr[T \in \mathsf{Close}] \leq e^{\epsilon} \Pr[T' \in \mathsf{Close}] + \delta'$$

We want to require that $\delta > \Pr[T \in \mathsf{Far}]$. We again separate $\Pr[T \in \mathsf{Far}]$ by

considering each coordinate separately and taking a union bound.

$$
\begin{aligned}
\delta' &\geq \Pr[T \in \mathsf{Far}] \\
&\geq \sum_j \Pr[T_j \notin [a_j + \tilde{w}_j, m - 1 - a_j + \tilde{w}_j]] \\
&\geq \sum_j (\Pr[T_j \leq a_j + \tilde{w}_j] + \Pr[T_j \geq m - 1 - a_j + \tilde{w}_j]) \\
&\geq \sum_j (cdf_{m-1}(a_j + \tilde{w}_j - c_j) + [1 - cdf_{m-1}(m - 1 - a_j + \tilde{w}_j - c_j)]) \\
&\geq \sum_j (cdf_{m-1}(a_j + \tilde{w}_j - c_j) + cdf_{m-1}((m-1) - (m - 1 - a_j + \tilde{w}_j - c_j))) \\
&\geq \sum_j (cdf_{m-1}(a_j + \tilde{w}_j - c_j) + cdf_{m-1}(a_j - \tilde{w}_j + c_j)) \\
&\geq \sum_j (cdf_{m-1}(a_j + \tilde{w}_j - w_{1,j}) + cdf_{m-1}(a_j - \tilde{w}_j + w_{2,j})) \\
&\geq 2 \sum_j cdf_{m-1}(a_j + .5 w_j)
\end{aligned}
$$

With this condition on δ' in place, we can reduce our condition to

$$
\begin{aligned}
\Pr[T \in \mathsf{Close}] &\leq e^{\epsilon} \Pr[T' \in \mathsf{Close}] \\
\Pr[T = y] &\leq e^{\epsilon} \Pr[T' = y] \text{ for all } y \in \mathsf{Close} \\
\frac{\Pr[T = y]}{\Pr[T' = y]} &\leq e^{\epsilon} \text{ for all } y \in \mathsf{Close} \\
\frac{\prod_j \Pr[T_j = y_j]}{\prod_j \Pr[T'_j = y_j]} &\leq e^{\epsilon} \text{ for all } y \in \mathsf{Close} \\
\prod_j \frac{\Pr[T_j = y_j]}{\Pr[T'_j = y_j]} &\leq e^{\epsilon} \text{ for all } y \in \mathsf{Close} \\
\prod_j \frac{pdf_{m-1}(y_j - c)}{pdf_{m-1}(y_j - \tilde{w}_j)} &\leq e^{\epsilon} \text{ for all } y \in \mathsf{Close}
\end{aligned}
$$

The worst case (for $y \in \mathsf{Close}$) occurs when $y_j = a_j + \tilde{w}_j$ and $c_j = w_{1,j}$ or when $y_j = m - 1 - a_j + \tilde{w}_j$ and $c_j = w_{2,j}$. Again, these are equivalent worst cases at opposite tails of the distribution. That means we must take ϵ such that

$$\prod_j \frac{pdf_{m-1}(a_j + \tilde{w}_j - w_{1,j})}{pdf_{m-1}(a_j + \tilde{w}_j - \tilde{w}_j)} \leq e^{\epsilon}$$

$$\prod_j \frac{pdf_{m-1}(a_j + .5w_j)}{pdf_{m-1}(a_j)} \leq e^{\epsilon}$$

We now know that the second privacy condition, Equation 6.30, is satisfied with

$$\epsilon = \sum_j \ln\left(\frac{pdf_{m-1}(a_j + .5w_j)}{pdf_{m-1}(a_j)}\right)$$

$$\delta' = 2\sum_j (cdf_{m-1}(a_j + .5w_j)).$$

Step 4: Comparing these two sets of values for ϵ and δ', we find the first ϵ and the second δ' values to be the limiting cases. Taking these gives us

$$\epsilon = \sum_j \ln\left(\frac{pdf_{m-1}(a_j)}{pdf_{m-1}(a_j - .5w_j)}\right)$$

$$\delta' = 2\sum_j (cdf_{m-1}(a_j + .5w_j).$$

We then need to get a final δ value by adding (a bound on) $(1 + e^{\epsilon})\Pr[\mathsf{fail}]$ to δ'. $\Pr[\mathsf{fail}]$ is the probability that fewer than m rows will remain random in the first run of the algorithm for Z. In other words, we're looking to bound the tail of the

binomial distribution. We use Hoeffding's inequality which gives

$$\begin{aligned}\Pr[\mathsf{fail}] &\leq e^{-2(nh-nhr)^2/n} \\ &\leq e^{-2(nh)^2(1-r)^2/n} \\ &\leq e^{-2nh^2(1-r)^2}.\end{aligned}$$

We do not have a formal optimization of r, but in general believe the best choice is the highest possible value that keeps the bound on $\Pr[\mathsf{fail}]$, given above, small enough that it does not contribute significantly to δ when compared to the other term.

Combining the above facts gives us the following final values.

$$\epsilon = \sum_j \ln\left(\frac{pdf_{\lceil rhn\rceil-1}(a_j)}{pdf_{\lceil rhn\rceil-1}(a_j - .5w_j)}\right)$$

$$\delta = 2\sum_j (cdf_{\lceil rhn\rceil-1}(a_j + .5w_j)) + (1+e^{\epsilon})e^{-2nh^2(1-r)^2}$$

□

Corollary 6.2 *Let $\mathcal{D}$ be the generating distribution that outputs a database X of n rows each chosen i.i.d. from a distribution over d-tuples of real numbers, with $X_{i,j}$ representing the j^{th} coordinate of the i^{th} row. Let the probability density function $p(y)$ contain a $d+1$-dimensional prism of volume v. That is, let prism $P = [s_1, t_1] \times \ldots \times [s_d, t_d]$ such that $p(y) \geq h$ for all $y \in P$ and some $h > 0$, and let $v = h\prod_i (t_i - s_i)$. Let $w_{1,j} = \inf(\mathsf{Supp}(X_{i,j}))$ and $w_{2,j} = \sup(\mathsf{Supp}(X_{i,j}))$ be the lower and upper bounds*

of j^{th} component of this distribution, with $w_j = w_{2,j} - w_{1,j}$ the width. Let $\Delta = \{\mathcal{D}\}$ (with empty/constant auxiliary information). Let F be the mechanism that outputs the (d-dimensional) exact sum of all rows of the database. F is $(\epsilon, \delta, \Delta)$-DDP with

$$\epsilon = \sum_j \ln\left(\frac{pdf_{\lceil rvn\rceil-1}(a_j)}{pdf_{\lceil rvn\rceil-1}\left(a_j - \frac{.5w_j}{t_j - s_j}\right)}\right)$$
$$\delta = 2\sum_j \left(cdf_{\lceil rvn\rceil-1}\left(a_j + \frac{.5w_j}{t_j - s_j}\right)\right) + (1 + e^{\epsilon})e^{-2nv^2(1-r)^2}$$

for any choice of values $a_j \in [0, n/2]$ and any $r \in [0, 1]$.

Proof: We first take the original data and apply linear transforms to to each coordinate so that each interval $[s_j, t_j]$ is mapped to $[0, 1]$. This makes the distribution satisfy the conditions of Theorem 6.7, with v in place of Theorem 6.7's h and $w_j/(t_j - s_j)$ in place of Theorem 6.7's w_j. As a result, we know that releasing the sum of this modified database is private. Theorem 5.3 then tells us that we can apply post-processing to this output, in particular inverting the aforementioned linear transforms, and output the result. This allows the true some of the original data to be output with privacy. □

Corollary 6.3 *Corollary 6.2 continues to hold when Δ consists of many distributions of the type discussed, with ϵ and δ values equal to the maximum value of those required by each distribution individually, and when Δ is further expanded to include convex combinations of such distributions (and therefore no longer independent rows). Furthermore, if Z is changed so that $Z = f(X')$ for some f, where $X' \subset X$ and $|X'| = \nu$, then the result still holds, but with n replaced by $n - \nu$.*

Proof: The ability to add distributions to Δ while taking the maximum of the ϵ and δ values that hold for each is a simple consequence of the transitivity of inequalities. Including convex combinations follows from Theorem 5.4. To include the auxiliary information, consider first the case where $Z = X'$, with the full subset of the database released. The proof would proceed as before, with the sum of X' added to the known sum t that is disregarded and the remaining number of rows being $n - \nu$ (before being reduced further by the added auxiliary information used in the proof). Once this is clear, the reduction of auxiliary information from the full subset X' to some function of it $f(X')$ is automatic from Theorem 5.5. □

It is worth emphasizing several things about the resulting general version of the theorem. First, it allows the release of truly new information. Some other definitions (e.g., noiseless privacy) generally require that rows be chosen independently from a known distribution in order for information like a sum/average to be released. This means that the query can never accomplish the most common goal, which is to estimate the average of the underlying population from which the database is sampled, or to give some indication of what type of sample the database represents, because such an answer is only useful in a situation where the average of the distribution is not already known. The end result applies for data that are, for example, known to be from an approximately normal distribution of unknown mean.

Second, the model of data distribution and auxiliary information here is quite reasonable. In the prior result on histograms, we used a very particular distribution. The goal there was to model exactly a common, realistic scenario. Here the goal is different. We assume only the existence of a prism of the sort described, a minimal

requirement on a continuous distribution. Furthermore, one can safely bound ϵ and δ if one is comfortable assuming lower bounds on the size of this prism. No additional information about the distribution is needed.

Furthermore, the model of auxiliary information here is quite strong. All that is required is the assurance that a given number of rows are entirely unknown to the adversary. Of course, in practice this number might be quite large in some situations, but the nature of the assumption is quite reasonable. For example, let us return to the example of Facebook releasing an average age of its users. An adversary might have a huge amount of auxiliary information about the ages of many Facebook users, but some users will be missing from this information, or they might be bot-generated accounts unrelated to real people. For a variety of reasons Facebook can comfortably assume that for any adversary there are at least many thousands of accounts for which no information is known. This is the sort of conservative upper bound that we advocate using. It is not attempting to precisely model the adversary's information, but even without such an attempt, the uncertainty assumed is sufficient in some cases for privacy without any need to add noise.

6.3.4 Example Parameters

The expressions for ϵ and δ in the above theorems are messy and maybe hard to understand. It is tempting to replace the *pdf* and *cdf* of the Irwin-Hall distribution with that of the approximating Gaussian, which is an extremely good approximation. Unfortunately, this is not very effective. The expression for ϵ is

extremely messy and adds little in the way of clarity. The expression for δ is even less useful, since the cumulative density function cannot be expressed in closed form. Instead we try to give a better of what parameters can be obtained by calculating numerical values for several distributions on the original data.

We begin by considering data from a uniform distribution on $[0, 1]$, using Theorem 6.6. The table below shows some ϵ and δ values that can be achieved. We consider three database sizes, and for each size we list several choices of a to give a general idea of what the tradeoff between ϵ and δ looks like.[2]

n	a	ϵ	δ
100	40	.634	6.83×10^{-4}
100	37	.835	8.18×10^{-6}
1000	460	.243	1.31×10^{-5}
1000	450	.303	4.82×10^{-8}
10000	4870	.0782	6.95×10^{-6}
10000	4850	.0902	2.12×10^{-7}
10000	4830	.102	4.07×10^{-9}

Table 6.1: Concrete parameter value options for several database sizes when data is drawn from a uniform distribution (over any range).

Whether these values are acceptable is highly dependent on the context being considered. In small, early-stage clinical trials with very few subjects, these results allow privacy only with horribly high parameters. For huge datasets based on website traffic, extremely good privacy can be achieved. (The Netflix dataset, for example, had millions of records from hundreds of thousands of individuals.)

[2]The calculations can become computationally non-trivial. We note that in general one can obtain several significant figures of accuracy by approximating the Irwin-Hall distribution by the Gaussian distribution.

We also want to show an example with non-uniform rows, so we consider rows generated according to a Gaussian distribution. No privacy is possible when row values are potentially infinite, so we use a Gaussian that is truncated at 2.5 standard deviations, meaning that values above or below 2.5 standard deviations from the mean are recorded as having a value of exactly 2.5 standard deviations above/below. This "top-coding" is a commonly used technique in data collection/analysis. (We also assume a standard Gaussian with mean of 0 and variance of 1, though this does not affect the resulting ϵ and δ values.)

When analyzing the privacy of sums in a Gaussian-distributed database, we have a choice of what interval to use for $[s, t]$. Remember that $[s, t]$ (and implicitly h) as well as r and a are all parameters of the analysis, not the underlying algorithm. They collectively represent a tradeoff between ϵ and δ, but for any value of ϵ or δ is desired, there is a "correct" choice of these parameters that minimizes the other value. Unfortunately, we do not have a closed-form solution for the optimal values given a desired ϵ or δ. Nevertheless, some basic rules can be found. It is, for example, always optimal to pick the largest h possible for the given $[s, t]$. In general, we want the area of the contained rectangle $(t - s)h$ to be large, and also want the width $t - s$ to be large. These goals can be contradictory, but in the case of the Gaussian distribution, we clearly want to be considering $[s, t]$ symmetric around zero and h equal to the probability density function at the ends. For that reason, we always pick $s = -t$ and list only t in the table below. h and w are determined automatically by the choice of t, but we list them to give an indication of the sort of values that can be achieved for a realistic distribution.

n	t	h	w	a	r	ϵ	δ
1000	1	.484	2.5	175	.83	.986	2.77×10^{-5}
1000	1.5	.389	1.67	125	.78	.904	7.43×10^{-7}
1000	1.5	.389	1.67	130	.79	.775	1.24×10^{-5}
1000	2	.216	1.25	56	.66	.829	4.73×10^{-5}
10000	1.5	.389	1.67	1725	.93	.227	3.69×10^{-6}
10000	1.5	.389	1.67	1690	.92	.273	2.89×10^{-8}

Table 6.2: Concrete parameter value options for several database sizes. In all cases the underlying distribution on database rows is a standard Gaussian distribution truncated at -2.5 and 2.5.

We note also that similar analysis to what we present above can be done with a variety of alterations to achieve better results for particular values. In particular, instead of using a uniform distribution (i.e., a rectangle-shaped probability density function) for the randomness-providing non-disclosed rows, one can pick another distribution. A symmetric triangle, for example, is the sum of two uniform distributions and can therefore be used with almost identical analysis. For peaked distributions, this could have greater area than the rectangle we use here. One could also use truncated Gaussian or Laplacian distributions to make better use of the full randomness available. One could also include multiple disjoint rectangles in the analysis, with Z either disclosing a row's value or saying which of the two (or more) uniform distributions the point came from. These options all require more complex analysis, but no fundamentally new ideas, and the results could provide substantial (constant-factor) improvements in ϵ and δ.

6.4 Linear Regression

Linear regression, also called ordinary least squares regression, is a commonly used method in statistics and machine learning. Here the data consists of data points, with a row x_i consisting of two values, g_i and h_i. g_i is a vector of dimension d consisting of real numbers measuring variables thought to offer possible predictive value. h_i is a scalar measurement of the variable that is being predicted. It is assumed that there is a vector β such that $h_i = g_i \cdot \beta + \mathsf{err}_i$ where err_i is random (hopefully small) error term.

To find the best-fitting value of β, there must be an accepted measurement of closeness between the predicted and actual values of h_i. In this case, that measurement is the square of the distance, so $\hat{\beta}$, our estimate for β, is the value that minimizes $\sum_i (h_i - g_i \cdot \hat{\beta})^2$. This quantity is uniquely minimized with [35]

$$\hat{\beta} = \left(\frac{1}{n} \sum_i ||g_i||^2 \right)^{-1} \times \frac{1}{n} \sum_i g_i h_i.$$

The expression above is important for our purposes because it shows that $\hat{\beta}$ can be computed using only a series of averages. In particular, one need only know the average value over all rows of $||g_i||^2$ and of each of d values of $g_{i,j} h_i$ (where $g_{i,j}$ is the j^{th} element of g_i). This is a set of $d + 1$ averages. Crucially, there are also $d + 1$ variables per line in the input (h_i and the d coordinates of g_i). The database rows can be thought of as containing, instead of the $d + 1$ values they actually contain, the $d + 1$ *implied* values whose averages are needed. In general,

because the number of values does not change, the resulting distribution on the implied values will satisfy the conditions of Corollary 6.2. The values of ϵ and δ, however, are highly dependent on the exact distribution. Nevertheless, this means that for a wide variety of distributions (that choose rows independently or are convex combinations of such distributions) we can give an exact linear regression output while maintaining privacy (and with ϵ and δ that approach zero as n grows).

There are, however, some caveats that should be emphasized. The result does not hold if the database rows do not each contain $d + 1$ degrees of freedom. If, for example, the dependent variable h_i is *perfectly* predicted (meaning $\mathsf{err}_i = 0$) by g_i, then the support of X_i is a d-dimensional surface in $\mathbb{R}^d$, meaning that there is no contained prism with positive volume. This also applies if two of the values in g_i are perfectly correlated. (This can be the case if, for example, a user wants to include the square of one of the variables as a way of modeling a potential quadratic relationship to h_i.) Perhaps most importantly, this excludes regressions using the practice of setting the first coordinate of g_i to a constant value of 1 to allow a predicting plane that does not pass through the origin. (Unfortunately, making one variable in the actual data constant or a function of the others does not in general make one of the $d + 1$ needed means constant or a function of the others.)

6.4.1 Simple Linear Regression

To show how one might deal with the limitations on releasing exact linear regression outputs, we consider the case of simple linear regression. This is linear

regression as discussed before, with two important changes. The first is that there is only one independent value g_i (no longer a vector). The second is that a constant term is included. As a result, we are assuming that the data is generated by $h_i = \alpha + g_i\beta + \mathsf{err}_i$, and we are trying to approximate α and β. The estimates $\hat{\alpha}$ and $\hat{\beta}$ that achieve the least squared error for a given sample are given by

$$\hat{\beta} = \frac{\overline{gh} - \overline{g}\,\overline{h}}{\overline{g^2} - \overline{g}^2}$$

$$\hat{\alpha} = \overline{h} - \beta\overline{g}$$

where, for example, $\overline{g}$ is the mean of all g_i values [35].

As expected, the inclusion of the constant term means that we need to compute more means than we have degrees of freedom in a row of the original data. Ideally we would want means of the vector $(g_i, h_i, g_i h_i, g_i^2)$, but any distribution on (g_i, h_i) would induce a distribution whose support is a 2-dimensional surface in 4-dimensional space and therefore has zero volume. To solve this problem, we split the database in two, and calculate two means each based on each of the two parts. To show that this is possible, we need the following theorem.

Theorem 6.8 *Say that database X is divided into two parts, so $X = (X', X'')$, and say that a mechanism F consists of separate mechanisms being run on X' and X'', so $F(X) = (F'(X'), F''(X''))$. Let the distribution $\mathcal{D}$ be one that generates each row independently from a given distribution. Then, if F' is (ϵ', δ')-DDP, and F'' is (ϵ'', δ'')-DDP, then F is $(\max(\epsilon', \epsilon''), \max(\delta', \delta''))$-DDP.*

Proof: Let $p(\cdot)$, $p'(\cdot)$ and $p''(\cdot)$ be the probability density functions of $F(X)$, $F'(X')$ and $F''(X'')$ respectively. Because X and X' are independent, we have $p((s_1, s_2)) = p'(s_1)p''(s_2)$. We define Sim to behave like F on the half of the database that is not missing a row, and on the other half behave like the simulator of F' or F''.

There are two probability inequalities that must be satisfied. Both proceed nearly identically, so we show the following requirement (for all sets S and all row indices i):

$$\Pr[F(X) \in S] \leq e^{\epsilon} \Pr[\mathsf{Sim}(X_{-i} \in S] + \delta$$

We begin with $Pr[F(X) \in S]$ and manipulate it to get the needed upper bound. We use the notation $S_{s_2} = \{s_1 \mid (s_1, s_2) \in S\}$ and $S_2 = \{s_2 \mid (s_1, s_2) \in S \text{ for some } s_1\}$. We assume that the missing row X_i is in X'.

$$\begin{aligned}
Pr[F(X) \in S] &= \iint_S p((s_1, s_2))\, ds_1\, ds_2 \\
&= \iint_S p'(s_1)p''(s_2)\, ds_1\, ds_2 \\
&= \int_{S_2} \int_{S_{s_2}} p'(s_1)p''(s_2)\, ds_1\, ds_2 \\
&= \int_{S_2} p''(s_2) \left[\int_{S_{s_2}} p'(s_1)\, ds_1 \right] ds_2 \\
&= \int_{S_2} p''(s_2) \Pr[F'(X') \in S_{s_2}]\, ds_2 \\
&\leq \int_{S_2} p''(s_2)(e^{\epsilon'} \Pr[\mathsf{Sim}'(X'_{-i}) \in S_{s_2}] + \delta')\, ds_2 \\
&\leq \delta' + e^{\epsilon'} \int_{S_2} p''(s_2) \Pr[\mathsf{Sim}'(X'_{-i}) \in S_{s_2}]\, ds_2 \\
&\leq \delta' + e^{\epsilon'} \Pr[\mathsf{Sim}(X) \in S]
\end{aligned}$$

A mirrored analysis reversing the roles of each half of X shows the same result when the missing row is in X'' (except that the parameters are ϵ'' and δ''). Similar analysis also shows the other needed inequality. □

We now find numerical estimates of what ϵ and δ values can be achieved. This will depend on the distribution from which the rows are drawn. Here we pick a simple example and use g_i and h_i both drawn from $[-1, 1]$ (so the true values of α and β are both zero). We first show privacy parameters for releasing the means of g_i and h_i.

n	a_1, a_2	r	ϵ	δ
1000	400	.9	.669	2.05×10^{-8}
5000	2270	.955	.295	1.14×10^{-8}
20000	9500	.976	.160	4.42×10^{-10}

Table 6.3: Concrete parameter value options for calculation of $(\overline{g}, \overline{h})$ when g_i and h_i both come from uniform distributions.

The rest of the database is used in computing $(\overline{g^2}, \overline{gh})$. To do this, we first introduce a change of variables, with $(j_i, k_i) = (g_i^2, g_i h_i)$ and J_i and K_i the random variables that take on j_i and k_i as specific values. We must determine the probability density function $p(j, k)$ for (J_i, K_i). We first calculate the probability density function $p_J(j)$ for J_i (for $0 \leq j \leq 1$). Note that the cumulative density function is $\sqrt{j}$. That means the probability density function of j is the derivative, $p_J(j) = \frac{1}{2\sqrt{j}}$. For any fixed value c of J_i, the integral of $p(j, k)$ along the line it defines ($j = c$) must be proportional to $\frac{1}{2\sqrt{j}}$. Furthermore, p will be constant along that line for k in the range $[-\sqrt{j}, \sqrt{j}]$ and 0 elsewhere. To get a value of $\frac{1}{2\sqrt{j}}$ when integrating a

constant function along an interval of length $2\sqrt{j}$, the value must be $\frac{1}{2\sqrt{j}*2\sqrt{j}} = \frac{1}{4j}$

This means we have

$$p(j,k) \propto \begin{cases} \frac{1}{4j} & : k \in [-\sqrt{j}, \sqrt{j}] \\ 0 & : k \notin [-\sqrt{j}, \sqrt{j}] \end{cases}$$

Using the above probability density function, we now evaluate the ϵ and δ values achieved for a calculation of $(\overline{g^2}, \overline{gh})$ on a database of various sizes. We apply Corollary 6.2 using $t_1 = 1$ and $t_2 = .6$, which naturally leads to $s_1 = .36$ and $s_2 = -.6$. The height is limited by the height at $(1, .6)$, which is .25. As a result, the volume is .192. $w_1 = 1$ and $w_2 = 2$.

n	a_1, a_2	r	ϵ	δ
7500	535	.82	.914	9.30×10^{-8}
10000	740	.84	.801	3.64×10^{-8}
14000	1082	.865	.676	3.02×10^{-8}
70000	6113	.938	.293	1.35×10^{-8}

Table 6.4: Concrete parameter value options for calculation of $(\overline{x}, \overline{y})$ when x and y both come from uniform distributions.

Combining the values in these two tables, we can arrive at values of the privacy parameters achievable at various database sizes. For example, we can have $\epsilon = .295$ and $\delta = 1.35 \times 10^{-8}$ with 75,000 total data points (5,000 of which are used for calculating the first pair of values). We can therefore get private values without the addition of any noise. However, the components of the regression coefficients are now being calculated over only part of the database, meaning that the answer is in some sense still inexact (though because it is deterministic, it might be easier to analyze and work with in some ways).

Chapter 7: Conclusion

In this dissertation we have explored two potential ways to weaken the differential privacy framework. In Chapter 4 we discussed *computational differential privacy*, introduced by Mironov et al. [56]. We believe the weakening is compelling — that is, it is indeed a weakening that does not lose the fundamental guarantee of privacy protection. However, we find it unlikely that this definition will lead to more accurate query output in the standard client-server setting, and we show two results to this effect.

In Chapter 5 we move to our own proposed weakening, *coupled-worlds privacy*, that takes into account the uncertainty inherent in the database, providing an alternative source of randomness. It assumes the database owner can place some reasonable upper bounds on what an adversary might already know about the data. This introduces a new source of potential error — the owner could be wrong about those upper bounds — but assuming no such error is made, we show that this definition still protects privacy. We also show a variety of useful properties and discuss several instantiations that capture a variety of particular privacy notions.

In Chapter 6 we then discuss *distributional differential privacy*, a particular instantiation of coupled-worlds privacy in more detail. We show that it can indeed

be used to output more accurate (often perfectly accurate) query output than differential privacy. In particular we discuss MAP estimators, histograms, sums and linear regression. We use both distributions meant to be precise models of realistic data generation (e.g., sampling distributions for histograms) and more general distribution families meant to provide conservative assumptions that would cover a wide variety of possible "true" distributions (e.g., in analyzing sums).

Open questions. A wide variety of questions remain open in this area. Many queries have not been studied. In particular, we believe that a variety of machine learning algorithms and statistical tools (e.g., hypothesis testing) provide promise. Even queries analyzed here could see improved results. While the mechanism itself generally cannot be improved (because most provide perfect accuracy), the analysis could provide better parameters or a wider class of acceptable distributions. In particular, under more specific assumptions about data generation (e.g., data drawn from normal distributions in each variable) linear regression output could probably be released with better privacy parameters. We also only conduct analysis under one of many possible instantiations, and all possible queries under other instantiations remain as open problems.

There is also substantial work to be done in analyzing the definition itself. We prove a variety of useful properties, but we suspect many more exist. Most importantly, our composition theorem is awkward and hard to use. This is probably the most important issue standing in the way of practical use of the definition. We do not have a counterexample that shows the assumptions in our composition theorem

are all necessary. A more widely applicable composition theorem, even if limited to certain instantiations or certain classes of data distributions, would be extremely useful.

Of course there are also a huge number of open questions in the wider field, with active research in differential privacy turning up new private mechanisms all the time. In addition to finding such mechanisms, there is a need to increase the ease of use of such mechanisms so that they become more frequently used by researchers. Only then can they replace the less reliable methods that are still too common outside the computer science community.

Bibliography

[1] Nabil Adam and John Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21(4):515–556, 1989.

[2] Michael Barbaro and Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749. The New York Times, August 9, 2006.

[3] Raef Bassily, Adam Groce, Jonathan Katz, and Adam Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *54th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2013.

[4] Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. In *7th Theory of Cryptography Conference (TCC)*, LNCS, pages 437–454. Springer, 2010.

[5] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *4th Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 97–110. ACM, 2013.

[6] Raghav Bhaskar, Abhishek Bhowmick, Vipul Goyal, Srivatsan Laxman, and Abhradeep Thakurta. Noiseless database privacy. In *Advances in Cryptology – Asiacrypt 2011*, pages 215–232, 2011.

[7] Abhishek Bhowmick and Cynthia Dwork. Natural differential privacy. In *DIMACS Workshop on Recent Work on Differential Privacy across Computer Science*, 2012.

[8] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical Privacy: The SuLQ Framework. In *24th Symposium on Principles of Database Systems*, pages 128–138. ACM, 2005.

[9] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *40th Annual Symposium on Theory of Computing (STOC)*, pages 609–618. ACM, 2008.

[10] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *46th Annual Symposium on Theory of Computing (STOC)*, pages 1–10. ACM, 2014.

[11] Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. *Journal of Machine Learning Research – Proceedings Track*, 19:155–186, 2011.

[12] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 289–296, 2008.

[13] Andrew Chin and Anne Klinefelter. Differential privacy as a response to the reidentification threat: The Facebook advertiser case study. *North Carolina Law Review*, 90:1417, 2011.

[14] Francis Y. Chin and Gultekin Özsoyoğlu. Statistical database design. *ACM Transactions on Database Systems*, 6:113–139, 1981.

[15] Francis Y. Chin and Gultekin Özsoyoğlu. Auditing and inference control in statistical databases. *IEEE Transactions on Software Engineering*, SE-8(6):574–582, 1982.

[16] Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:429–444, 1977.

[17] Anindya De. Lower bounds in differential privacy. In *9th Theory of Cryptography Conference (TCC)*, LNCS, pages 321–338. Springer, 2012.

[18] Dorothy E. Denning. Secure statistical databases with random sample queries. *ACM Transactions in Database Systems*, 5:291–315, 1980.

[19] David Dobkin, Anita K. Jones, and Richard J. Lipton. Secure databases: Protection against user influence. *ACM Transactions in Database Systems*, 4:97–106, 1979.

[20] Josep Domingo-Ferrer and Vicenç Torra. A critique of k-anonymity and some of its enhancements. In *3rd International Conference on Availability, Reliability and Security (ARES)*, pages 990–993. IEEE, 2008.

[21] Yitao Duan. Privacy without noise. In *18th Conference on Information and Knowledge Management (CIKM)*, pages 1517–1520. ACM, 2009.

[22] Cynthia Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1–12. Springer, 2006.

[23] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology – Eurocrypt 2006*, pages 486–503. Springer, 2006.

[24] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *41st Annual Symposium on Theory of Computing (STOC)*, pages 371–380. ACM, 2009.

[25] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *3rd Theory of Cryptography Conference (TCC)*, LNCS, pages 265–284. Springer, 2006.

[26] Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 2010.

[27] Ivan P. Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67(337):7–18, 1972.

[28] Theodore D. Friedman and Lance J. Hoffman. Towards a fail-safe approach to secure databases. In *1st Symposium on Security and Privacy (SP)*, pages 18–21. IEEE, 1980.

[29] Johannes Gehrke, Michael Hay, Edward Lui, and Rafael Pass. Crowd-blending privacy. In *Advances in Cryptology – Crypto 2012*, pages 479–496, 2012.

[30] Johannes Gehrke, Edward Lui, and Rafael Pass. Towards privacy for social networks: A zero-knowledge based definition of privacy. In *8th Theory of Cryptography Conference (TCC)*, LNCS, pages 432–449. Springer, 2011.

[31] Rosario Gennaro, Yael Gertner, Jonathan Katz, and Luca Trevisan. Bounds on the efficiency of generic cryptographic constructions. *SIAM Journal on Computing*, 35(1):217–246, 2005.

[32] Adam Groce, Jonathan Katz, and Arkady Yerukhimovich. Limits of computational differential privacy in the client/server setting. In *8th Theory of Cryptography Conference (TCC)*, LNCS, pages 417–431. Springer, 2011.

[33] Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Random differential privacy. *Journal of Privacy and Confidentiality*, 4(2):43–59, 2012.

[34] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *42nd Annual Symposium on Theory of Computing (STOC)*, pages 705–714. ACM, 2010.

[35] Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

[36] Lance J. Hoffman and W.F. Miller. Getting a personal dossier from a statistical data bank. *Datamation*, 16(5):74–75, 1970.

[37] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of

DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genetics*, 4, 2008.

[38] Neil Hunt. Netflix Prize Update – Netflix US and Canada Blog. `http://blog.netflix.com/2010/03/this-is-neil-hunt-chief-product-officer.html`, March 12, 2010.

[39] Russell Impagliazzo and Steven Rudich. Limits on the provable consequences of one-way permutations. In *21st Annual Symposium on Theory of Computing (STOC)*, pages 44–61. ACM, 1989.

[40] Inter-university Consortium for Political and Social Research. *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle*. ICPSR, Ann Arbor, MI, fifth edition, 2012.

[41] Inter-university Consortium for Political and Social Research. Confidentiality. `http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/index.html`, Accessed May 2014.

[42] Inter-university Consortium for Political and Social Research. Recommended informed consent language for data sharing. `http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/conf-language.html`, Accessed May 2014.

[43] Matthew Karnitschnig. AOL tech chief resigns over issue of released data – The Wall Street Journal. `http://online.wsj.com/news/articles/SB115618361010241207`, August 23, 2006.

[44] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *49th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 531–540. IEEE, 2008.

[45] Shiva Prasad Kasiviswanathan and Adam Smith. On the 'semantics' of differential privacy: A bayesian formulation. *Computing Research Repository (CoRR)*, arXiv:0803.39461 [cs.CR], 2013.

[46] Daniel Kifer and Ashwin Machanavajjhala. No Free Lunch in Data Privacy. In *SIGMOD*, pages 193–204, 2011.

[47] Daniel Kifer and Ashwin Machanavajjhala. A rigorous and customizable framework for privacy. In *31st Symposium on Principles of Database Systems*, pages 77–88, 2012.

[48] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *23rd International Conference on Data Engineering (ICDE)*, pages 106–115. IEEE, 2007.

[49] Chong K. Liew, Uinam J. Choi, and Chung J. Liew. A data distortion by probability distribution. *ACM Transactions on Database Systems*, 10:395–411, 1985.

[50] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *24th International Conference on Data Engineering (ICDE)*, pages 277–286. IEEE, 2008.

[51] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[52] Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *51st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 81–90. IEEE, 2010.

[53] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 94–103. IEEE, 2007.

[54] Frank D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *SIGMOD International Conference on Management of Data (SIGMOD)*, pages 19–30. ACM, 2009.

[55] Elinor Mills. AOL sued over Web search data release – CNET. `http://www.cnet.com/news/aol-sued-over-web-search-data-release/`, September 25, 2006.

[56] Ilya Mironov, Omkant Pandey, Omer Reingold, and Salil Vadhan. Computational differential privacy. In *Advances in Cryptology – Crypto 2009*, pages 126–142. Springer, 2009.

[57] Arvind Narayanan and Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets (How to Break Anonymity of the Netflix Prize Dataset). *IEEE Symposium on Security & Privacy*, pages 111–125, 2008.

[58] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *39th Annual Symposium on Theory of Computing (STOC)*, pages 75–84. ACM, 2007.

[59] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57(6), 2010.

[60] OkCupid. OkTrends blog. `http://blog.okcupid.com`.

[61] Gultekin Özsoyoğlu and Francis Y. Chin. Enhancing the security of statistical databases with a question-answering system and a kernel design. *IEEE Transactions of Software Engineering*, 8(3):223–234, 1982.

[62] H. Vincent Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, 1994.

[63] Ben Popken. AOL User 927, The Theatrical Production – Consumerist. `http://consumerist.com/2008/04/29/aol-user-927-the-theatrical-production/`, April 29, 2008.

[64] Omer Reingold. Occupy Database — Privacy is a Social Choice – Windows on Theory. `http://windowsontheory.org/2012/02/28/occupy-database-privacy-is-a-social-choice/`, February 28 2012.

[65] Omer Reingold, Luca Trevisan, and Salil Vadhan. Notions of reducibility between cryptographic primitives. In *1st Theory of Cryptography Conference (TCC)*, LNCS, pages 1–20. Springer, 2004.

[66] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.

[67] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, Technical report, SRI International, 1998.

[68] Jan Schlörer. Identification and retrieval of personal records from a statistical data bank. *Methods of Information in Medicine*, 14(1):7–13, 1975.

[69] Ryan Singel. Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims – WIRED. `http://www.wired.com/2009/12/netflix-privacy-lawsuit/`, December 17, 2009.

[70] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[71] Joseph F. Traub, Yechiam Yemini, and Henryk Woźniakowski. The statistical security of a statistical database. *ACM Transactions on Database Systems*, 9:672–679, 1984.

[72] US Code of Federal Regulations. Other requirements relating to uses and disclosures of protected health information. 45 CFR Subtitle A §164.514, 2002.

[73] Bimal Viswanath, Emre Kiciman, and Stefan Saroiu. Keeping information safe from social networking apps. In *Workshop on Online Social Networks (WOSN)*, pages 49–54. ACM, 2012.

[74] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60:63–69, 1965.

[75] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment (VLDB)*, 5(11):1364–1375, 2012.