Differential Privacy: An Introduction for Social Scientists (Rough Draft – Not for Circulation)

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

[1]Center for Research on Computation and Society, Harvard University.
{kobbi|mbun|tsteinke|salil}@seas.harvard.edu.
[2]Berkman Center for Internet & Society, Harvard University.
{awood|dobrien}@cyber.law.harvard.edu.
[3]School of Computing, University of Dundee. m.gaboardi@dundee.ac.uk.

October 18, 2015

Abstract

Using simple technical terms and limited mathematical formalism, we introduce *differential privacy* – a new privacy concept from the computer science literature – to a social science audience.

This is an excerpt of a larger document in progress that aims to provide social scientists with an overview of privacy-related topics relevant to the research use of information collected from individuals. The final document will cover the following topics: the importance of data privacy generally, the implications of privacy breaches, relevant laws and best practices, common de-identification methods and the strengths and weaknesses of such methods, re-identification risks, and differential privacy.

This work is the product of a working group of the *Privacy Tools for Sharing Research Data* project at Harvard University.[1] Our goal is to offer language that can be used to explain privacy-related topics to future users of the Dataverse[2] repository platform as they consider whether and how to use the privacy-enhancing tools being developed in the project. The working group discussions were led by Kobbi Nissim. Kobbi Nissim and Alexandra were the lead authors of this document. Working group members Mark Bun, David O'Brien, Marco Gaboardi, Kobbi Nissim, Thomas Steinke, Salil Vadhan, and Alexandra Wood contributed to the conception of the document and to the writing. We thank Caper Gooden and Daniel Muise for their valuable comments on an earlier version of this document.

This material is based upon work supported by the National Science Foundation under Grant No. 1237235 as well as the Sloan Foundation.

Keywords: differential privacy

[1]See http://privacytools.seas.harvard.edu/.
[2]See http://dataverse.org/.

1 Introduction

Differential privacy is a new privacy concept that has emerged in the computer science literature, in light of the accumulated evidence about weaknesses of traditional statistical disclosure limitation techniques, such as de-identification techniques. The failure of de-identification techniques to provide adequate privacy has demonstrated that commonly used privacy protection techniques may be subjected to attacks, such as record linkage attacks, devised after a privacy technique's deployment and use. Such attacks have also demonstrated the inability to remedy privacy breaches by "taking data back", as many copies of the de-identified data often exist and are accessible via the internet. These issues have highlighted the need for a privacy technology that is immune not only to linkage attacks, but to any potential attack, *including attacks that are currently unknown or unforeseen*. Furthermore, many attacks on de-identified data were successful due to the attackers' ability to combine the de-identified data with other available information, highlighting the need for a standard that provides meaningful privacy not only in a "standalone" setting, but also in combination with other information that may be available to attackers.

Differential privacy, presented in 2006, is the result of ongoing research to develop a privacy guarantee that provides robust protection even against unforeseen attacks. The concept is most readily applicable in contexts where some aggregate statistics are computed over a large collection of individual information. In the following sections, we provide an informal description of differential privacy. Using simple technical terms and without delving into the mathematical formalism, we discuss the definition of differential privacy, how it addresses privacy risks, how differentially private analyses are constructed, and how differential privacy can be used. We conclude with some advanced topics and pointers for further reading.

2 What is the differential privacy guarantee?

Consider a statistical procedure. The procedure receives as input information about individuals that is potentially privacy-sensitive. It performs some computation – an analysis of the data – and outputs the results of this computation. For example, it may compute the average income of a set of individuals, or perform regression analysis to estimate the correlation between a student's undergraduate GPA and her high school GPA and parents' income, or even apply a statistical disclosure limitation (SDL) procedure to aggregate or de-identify the data, with the goal of producing a sanitized version of the data that would be safe to share or disclose. Our notion of analysis (we will also use the term computation) is very broad and includes these and many other analyses. In short, we use this term to refer to any procedure for transforming the input data into some output (see Figure 1).

Intuitively, the requirement of differential privacy is that the output of an analysis should not reveal information about any specific individual. To illustrate how this requirement is formalized, consider the following scenario.

A sample of individuals was selected for to study the relationship between the financial conditions and medical outcomes in various American cities. The individuals are asked to answer a questionnaire and give information about where they live, their health, and their finances. John, who is in the sample, is aware of re-identification attacks that have been performed on de-identified data. He is worried that some sensitive information about him, such as his HIV status, or his yearly income, might be revealed by the analysis. If leaked, such information could affect his life

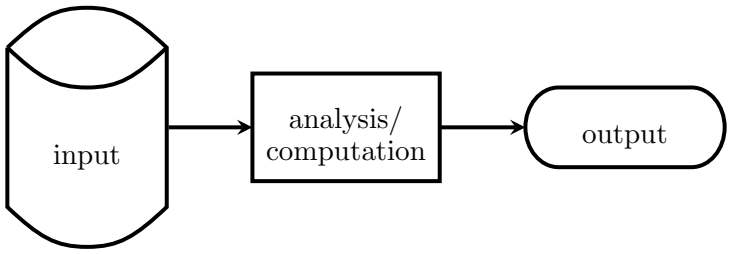

Figure 1: An analysis/computation

insurance premium, or the chance he will get a bank loan approved.

If the analysis is differentially private, then John is guaranteed that even though his information is used by the analysis, the outcome of the analysis would not disclose anything that is *specific to him*. To understand what this means, consider a thought experiment, which we refer to as John's privacy-ideal scenario and illustrate in Figure 2. Notice that John's information is "opted-out". In other words, in our thought experiment John's information is omitted from the input to the analysis, but the information of all other individuals is provided as input as usual. In this case, the outcome of the analysis cannot depend on John's specific information – because John's data is omitted, the outcome would remain the same even if John's information had been completely altered – and hence no information specific to John can be learned from the outcome. In John's privacy-ideal scenario, John's privacy is clearly preserved.

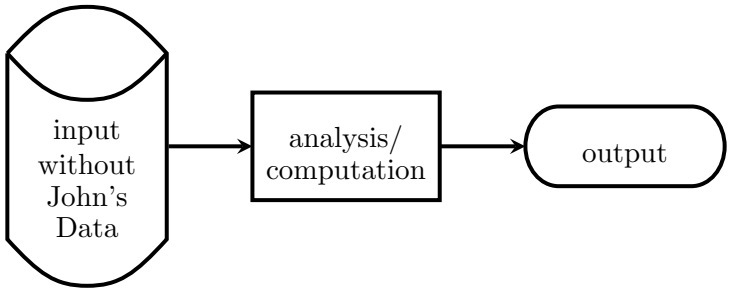

Figure 2: John's privacy-ideal scenario.

The real world scenario (as depicted in Figure 1) is, however, that John's information *is* included along with the information of others. Hence, there is a potential risk to his privacy, as sensitive information about him could be revealed in the analysis of the data. The guarantee of differential privacy is that the risk to John's privacy in the real world scenario would be essentially the same as in the privacy-ideal scenario. Hence, **what can be learned about John from a differentially private computation is (essentially) limited to what could be learned about him from everyone else's data *without him being included in the computation***. Crucially, this very

same guarantee is made not only with respect to John, but also to all other individuals contributing their information to the analysis!

A more precise description of this guarantee requires the use of formal mathematical language, and arguing formally about the properties of differential privacy requires the use of concepts and reasoning that are beyond the scope of this description. Rather than providing a full, precise definition, we give a few illustrative examples to discuss various aspects of differential privacy in a way we hope is more intuitive and accessible to a general audience.

2.1 What doesn't differential privacy protect?

Differential privacy aims to provide John with privacy protection that approximates his privacy-ideal scenario. Hence, we need to understand what is and what is not protected in John's privacy-ideal scenario. We will see that, even in the privacy-ideal scenario, information can be revealed about John that might embarrass him, harm his social status, or adversely affect his employability or insurability.

To better understand these risks, consider an observer, Alice, with prior knowledge of some information about John; say, Alice knows that John enjoys a lot of red wine. Suppose the survey reveals that there is a correlation between drinking red wine and a certain type of cancer. In John's privacy-ideal scenario, his information is omitted from the survey, and the analysis is performed over the information from all other individuals in the sample. Notice that even without John's information being included the analysis, Alice would learn from its output that he has a heightened cancer risk (just like other red wine drinkers). As a more extreme example, suppose Alice knows that John is a public school teacher and that all public school teachers receive the same standard salary (although she doesn't know what that standard salary is, as it may have recently been renegotiated). The survey results reveal how much other public school teachers earn. Thus Alice is able to exactly determine John's salary from the survey results in the privacy-ideal scenario, despite his data not being included in the survey. In both examples, in spite of not being included in the analysis, John may be adversely affected by the disclosure of a sensitive piece of information about him.

To summarize, John's privacy-ideal scenario and, thereby, differential privacy does not guarantee that *no* information about John can be revealed. It only guarantees that *no information specific to John* is revealed. In fact, any useful analysis carries a risk of revealing information about individuals, as demonstrated by the two examples above. We argue, however, that this kind of risk is unavoidable. In a world where data are collected and analyzed, the best privacy protection John could hope for is to opt out of allowing his data to be used in an analysis, i.e., enforce his privacy-ideal scenario.

2.2 What does differential privacy protect?

3 The privacy loss parameter

Let us revisit the privacy-ideal scenario. Consider the task of estimating the fraction of HIV+ people in the surveyed population. Ideally, the output of the analysis should remain exactly the same whether or not John participates in the survey. This seems a reasonable requirement, as John is only one of many people participating in the survey. However, privacy is not guaranteed only to John. To protect another individual's privacy as well, such as Gertrude's, we also need to consider

her privacy-ideal scenario. Hence, the outcome should remain the same when both her and John's inputs are omitted and so on with every other surveyed individual. Ultimately, we might reach the undesirable conclusion that the output of the analysis should remain the same when all inputs are omitted, and hence that the analysis must be useless!

To avoid this problem, differential privacy (Figure 3) only requires that the output of the analysis should remain *approximately* the same whether or not John participates in the survey. That is, differential privacy permits a slight deviation between the output of the analysis and that of each individual's privacy-ideal scenario. A parameter quantifies and limits the extent of this deviation. This parameter is usually denoted by the Greek letter ϵ (epsilon) and referred to as the privacy parameter, or, more accurately, the privacy loss parameter.[3] The parameter ϵ measures the effect of each individual's information on the output of the analysis. It can also be viewed as measuring the excess privacy risk to each individual, or how much privacy risk he or she could incur beyond the risk incurred within that individual's privacy-ideal scenario. Note that in Figure 3 we replaced John with a prototypical individual X to emphasize that the differential privacy guarantee is made simultaneously to *all* individuals in the sample.

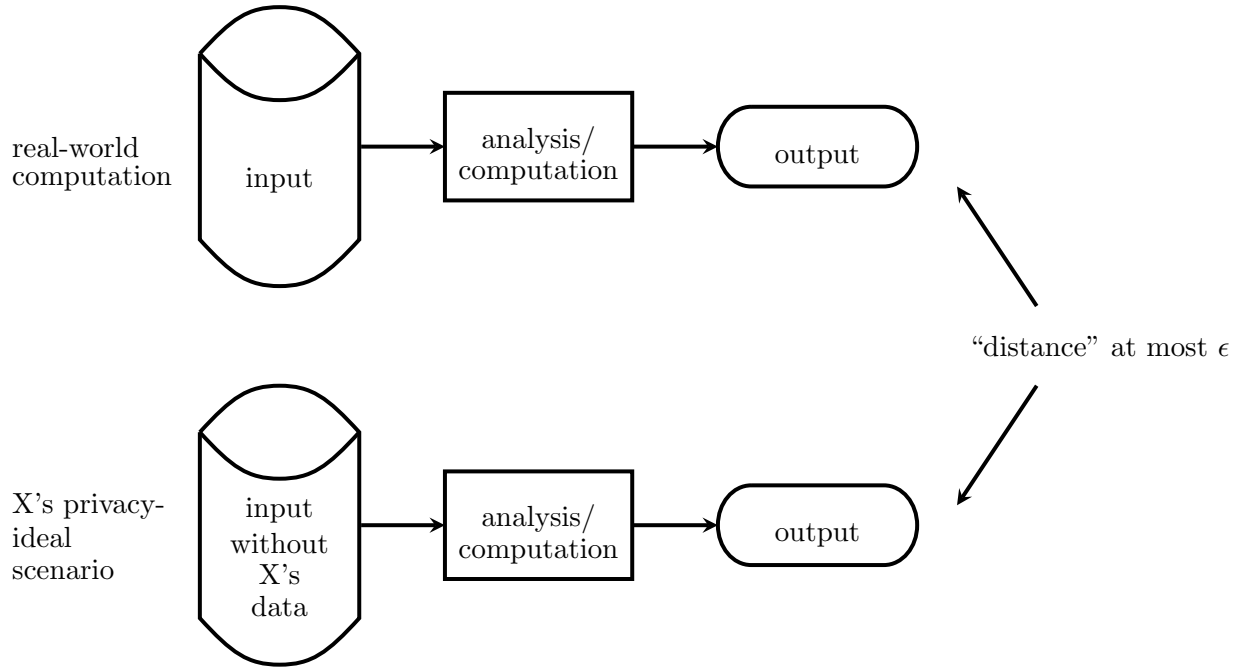

Figure 3: Differential Privacy. The distance requirement between the privacy-ideal scenario and real world computation should hold simultaneously for all individuals X whose data is in the input.

Controlling ϵ can be thought of as tuning the level of privacy protection required. A smaller ϵ means a smaller deviation between the real-world analysis and the privacy-ideal scenario, and

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

hence a stronger privacy guarantee. Note, however, that the choice of ϵ also affects the utility or accuracy that can be obtained from the analysis. For example, when ϵ is set to zero, the real-world differentially private analysis mimics the privacy-idea scenario of all individuals perfectly, but, as we argued in the first paragraph of this section, a simultaneous mimicking of the privacy-ideal scenarios of all individuals cannot provide any meaningful output. More generally, a smaller value of ϵ is associated with better privacy but lower accuracy. Hence, the value of ϵ should be chosen so as to allow a reasonable compromise between privacy and accuracy. As a rule of thumb, ϵ should be thought of as a small number, between approximately 1/1000 and 1.

We note that, for technical reasons, any analysis that is differentially private must be probabilistic in nature. The reader may be familiar with analyses performed using standard statistical software where the outcome is deterministic, meaning that executing the same analysis on the same data produces the same results. In contrast, executing a differentially private analysis several times on the same data can result in different answers. This is because differentially private analyses introduce some uncertainty into the computation in form of random noise. (See Section 6 for a discussion of how such analyses are constructed.)

We now discuss the effect of choosing the parameter ϵ more technically, and a reader who is coming across this concept for the first time may choose to skip to Section 3.1. For concreteness, we will use a value of 0.01 for ϵ.

Taking into account the probabilistic nature of differentially private analyses, we can now discuss how ϵ controls the effect of each individual's information on the outcome. We will use the notion of an event defined over the outcome of an analysis. For a concrete example, consider an analysis computing the fraction of HIV+ individuals in the surveyed population. The outcome of such an analysis is a number between 0 and 1. As we noted above, the differentially private analysis cannot simply output the exact fraction. Instead, the analysis outcome is noisy – it samples its outcome from a distribution of answers centered at the exact fraction (the accuracy of the analysis is related to the variance of this distribution). An event is simply a subset of the potential answers, for example, we can define the event:

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

Executing the analysis on its input, event E_1 would occur with some probability p. Similarly, given the input without John's data, event E_1 would occur with probability p'. The guarantee of differential privacy is that these probabilities are almost the same: the ratios p/p' and p'/p are both at most

$$1 + \epsilon = 1 + 0.01 = 1.01.$$

(Note: this analysis is approximate and holds only for small values of ϵ.)

More generally, differential privacy guarantees this bound not only for event E_1 but for every event defined over the outcome of the analysis, and the guarantee is made not only to John, but simultaneously to all the individuals whose information is used in the analysis.

Clearly, John is not directly concerned with the event E_1. John may be worried that his insurer will deny him coverage if it learns certain information about his health status via the differentially private mechanism. Notice that if John's insurer takes the outcome of the analysis into consideration when deciding whether to deny John's coverage, then the insurer's decision to deny coverage corresponds to an event defined over the outcome of the analysis. For instance, it may be that the insurer denies coverage when the outcome is between 0.1 and 0.2. Interestingly, we do not need to know how John's insurer reaches a decision (the decision may depend on multiple

factors, including what the insurer already knows about John and what the insurer learns from the differentially private mechanism). It is sufficient to notice that any such decision corresponds to an *event* over the output of the analysis, and hence it is guaranteed that the probability of John being denied coverage would not increase by a factor of more than $1 + \epsilon$ should John's information be included in the input to the mechanism (compared to the case where his information is not included).

If John believes that, without being included in the analysis, his probability of being denied insurance is at most $p = 5\%$, then having his information used in the differentially private computation can increase his probability being denied insurance to at most

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

3.1 Composition of differentially private analyses

It is unavoidable that the more analyses John's data is used in, the greater the risk to his privacy. The parameter ϵ can help us understand how this risk accumulates. Continuing the above example, let us suppose John's data is used in two differentially private analyses, each with privacy loss parameter $\epsilon = 0.01$. Above, we saw that John's probability of being denied insurance could increase from $p = 5\%$ to at most $p' = 5.05\%$ due to the inclusion of his information in one analysis. This value $p' = 5.05\%$ becomes John's baseline for the second analysis, and after the second analysis his probability of being denied insurance can rise to at most

$$p'' = p' \cdot (1 + \epsilon) = p \cdot (1 + \epsilon)^2 \approx p \cdot (1 + 2 \times \epsilon) = 5\% \times 1.02 = 5.1\%.$$

The exact analysis of how privacy loss accumulates is beyond the scope of this document and is captured by what researchers of differential privacy call composition theorems. These theorems provide formal bounds on the accumulated privacy risk from a collection of analyses given their privacy parameters.

The fact that the more John's information is used in analyses, the more risk there is to his privacy is not singular to differential privacy. It follows from fundamental properties of information that an increase in privacy risk must be incurred regardless of the privacy protection method in use. The distinguishing difference between differential privacy and other known measures of privacy is that, to date, it is the only protection of privacy for which the accumulated risk can be analysed and bounded.

4 How does differential privacy address privacy risks?

To illustrate the quantitative bounds on privacy risks that differential privacy can guarantee we now consider a concrete example. Gertrude is a 65-year-old woman who is considering whether or not to participate in a medical study. While Gertrude can see the potential benefits from the medical study, she is concerned that the information she reveals while participating in the study will affect her life insurance premium. However, for a medical study employing differential privacy, Gertrude can be assured that her participation will not change her life insurance premium by very much.

4.1 A baseline - Gertrude's privacy-ideal scenario

Based solely on her age and gender, Gertrude has a 1% chance of dying in the next year.[4] The value of her life insurance policy is $\$100,000$; therefore, a fair annual premium for her would be $\$1,000$, though the life insurance company would presumably charge more in order to make a profit. Gertrude is concerned that the tests she will undergo as part of the study will show that she is predisposed to suffer a stroke. This would mean she is significantly more likely to die in the coming year. If this information is made available to her insurance company, her premium will increase substantially.

If Gertrude opts not to participate in the study, her premium may increase regardless. For example, if a medical study concludes that coffee drinkers are more likely to suffer a stroke, Gertrude's insurance company may conclude that a 65-year-old female who drinks coffee has a 2% chance of dying in the next year. Because Gertrude is an avid coffee drinker, her "fair" premium would increase to $\$2,000$. This is the potential effect of an analysis we use as a basis for comparison, as this increase is unavoidable to Gertrude. Using the terminology of Section 2, this baseline corresponds to Gertrude's insurance premium in her privacy-ideal scenario.

4.2 Reasoning about Gertrude's risk

Suppose Gertrude does participate in the study, and the medical researchers conclude that, based on many tests, Gertrude has a 50% chance of dying from a stroke in the next year. If all of the research data were made public, her insurance company might increase her premium to over $\$50,000$.

Fortunately for Gertrude, this is not the case, as only a differentially private summary of the research data is released. How much might Gertrude's premium increase? If the medical study uses a value for the differential privacy parameter ϵ of 0.01, then the insurance company's estimate of the probability that Gertrude will die increases from 2% to at most

$$2\% \cdot (1 + 0.01) = 2.02\%.$$

Thus the fair premium can increase to no more than $\$2,020$, meaning the first-year cost in terms of insurance premium to Gertrude for participating is at most \$20.

More generally, we can use Table 1 to perform such a calculation. The top row shows the value of ϵ (in this case 0.01) used in the study. The left column shows the baseline probability (in this case 2%), which is the estimated probability of a given event assuming a given individual is not included in the study. The entry in the table then shows the worst-case probability (in this case 2.02%), which is the maximum possible estimated probability of the event if the individual is included in the study.

Note that Gertrude may decide that \$20 is too great a potential cost to warrant participation, and she may decide not to participate in the study with this value of ϵ and this level of risk. Alternatively, she may decide that it is worthwhile because perhaps she is paid more than \$20 to participate in the study or the information she learns from the study is valuable to her. The key is that differential privacy allows her to make an informed estimate of the costs of participation.

Note that the above calculation requires a lot of information. In particular, the 2% baseline is dependent on the outcome of the study, which Gertrude does not know when deciding to participate.

[4] See `http://www.ssa.gov/oact/STATS/table4c6.html`.

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

Table 1: Maximal change between posterior beliefs in Gertrude's privacy-ideal and real scenarios.

Fortunately, differential privacy provides guarantees for every event and every baseline value, as shown in the above table.

Lastly, note that Gertrude may experience other costs from participating in the study. For example her health insurance costs may also be affected and this would require a separate calculation.

5 Differential privacy and legal requirements

Legal standards refer to a less formal notion of risk than measured by differential privacy. Privacy laws generally focus on the ability to identify an individual's sensitive information in a release of records. While the ability to identify sensitive information is indeed a serious privacy breach, weaker forms of leakage, including the leakage of imprecise or incomplete individual information, may exist. Differential privacy provides protection with respect to the full range of such breaches.

For this reason, it seems that differential privacy provides stronger protection than privacy laws require. An examination of regulations such as the Family Educational Rights and Privacy Act (FERPA) and the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, and guidance from agencies on complying with their requirements, reveals that lawmakers seem to have had a significantly weaker notion of privacy attacks in mind than those captured and countered by the concept of differential privacy. For example, the HIPAA Privacy Rule permits health care providers to share health records that have been de-identified by removing certain fields such as names, addresses, telephone numbers, and Social Security numbers. This standard assumes a relatively weak attacker – one that is unlikely to successfully link the fields remaining in the data with information from other sources, such as voter registration lists, to identify individuals in the data. In fact, numerous re-identification demonstrations have shown it is possible to identify individuals in records that have been redacted according to this standard. Therefore, the standard seems to accept that some small number of re-identification attacks will be successful, and strikes a practical balance between mitigating re-identification risks and enabling beneficial uses of the data.

Because a data release that satisfies the definition of differential privacy provides robust protection against these types of linkage attacks plus other types of inferencing attacks, including those that are currently unknown or unforeseen, such an approach should be sufficient for complying with legal requirements. One needs to keep in mind that privacy definitions appearing in regulation do not follow the same definitional pattern and formalism of differential privacy, and additional research is needed to verify that differential privacy satisfies the requirements of regulations like FERPA and HIPAA.

6 How are differentially private analyses constructed?

The construction of differentially private analyses relies on the careful introduction of uncertainty in form of random noise. In the following section, we give a simple example of how this done. This explanation is somewhat technically involved, and the reader may choose instead to skip to Section 6.1 on a first read.

Consider computing an estimate of the number of HIV+ individuals in a sample of n individuals and denote by m the true number of HIV+ individuals in the sample. In a differentially private version of the computation, random noise Y, scaled so as to hide the contribution of a single individual, would be added to the count. The outcome of the analysis would then be

$$m' = m + Y.$$

One possibility is to sample the random noise Y from the Normal distribution with zero-mean and standard deviation $1/\epsilon$.[5] We get an immediate tradeoff between privacy and utility as the choice of ϵ affects the noise magnitude. The smaller ϵ is the more noise is added.

If a researcher uses the estimate m' for computing the true fraction p of HIV+ people in the population, then his computation would result in the estimate

$$p' = m'/n.$$

We note that there are two sources of error in estimating p. The sampling error would cause m to differ from the expected $p \cdot n$ by an amount of roughly

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

and the addition of noise would cause m' and m to differ by an amount of roughly

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

Overall, we get that the standard deviation of the estimate p' (hence the expected difference between p' and p) would be of magnitude roughly

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

which means that for a large enough sample size n the sampling error would dominate the noise added for privacy.

For more complex analyses, this simple noise addition technique is often sub-optimal in terms of accuracy, and the theoretical work on differentially private algorithms has identified many other noise introduction techniques that result in better accuracy guarantees.

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

6.1 What analyses can be performed with differential privacy?

We give a non-comprehensive list of analyses for which differentially private algorithms are known to exist:

- **Count queries:** The most basic statistical tool, a count query returns an estimate of the number of individual records in the data satisfying a specific predicate, e.g., the number of records corresponding to HIV+ individuals as in the example above. Differentially private answers to count queries can be obtained by the addition of random noise as discussed with counting HIV+ individuals above.

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

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

- **Linear regression:** An analysis of how a the value of a dependent variable varies as a function of one or more explanatory variables, assuming a linear model.

- **Clustering:** An exploratory data analysis technique where data points are grouped into clusters such that points in the same cluster are more similar to each other than to points in other clusters.

- **Classification:** Classification is the problem of categorizing data points into a set of categories, based on a training set of examples for which category membership is known.

7 Limits of differential privacy

7.1 Accuracy

A consequence of differential privacy is that random noise must be introduced into computations, and that this noise should be sufficiently large to hide the contribution of roughly any $1/\epsilon$ individuals. This means that differentially private computations are less accurate than the statistics one could directly compute on the data, or, put otherwise, carry a toll in the minimal sample size required. Much of the effort spent in studying differential privacy is focused on understanding and improving this tradeoff: how to obtain the maximum possible utility from data while preserving differential privacy.

In practice, the level of noise added to differentially private analysis means that little utility can be made of small- to moderately-sized datasets, and as a rule of thumb, almost no utility is expected from datasets containing $1/\epsilon$ or less records. For specific analyses, procedures for estimating the accuracy of an analysis based on properties of the collected data exist (such as those implemented in Harvard's Dataverse project, discussed below). These procedures take as input parameters the number of records, a value for ϵ, the range of numerical and categorical fields, etc. and produce bounds on the accuracy for a variety of statistical computations. Alternatively, a desired accuracy is given as input instead of ϵ, and the computation results in a value for ϵ that would afford this level of accuracy.

7.2 The "privacy budget"

We mentioned above (Section 3.1) that the composition of two differential privacy analyses results in risk that is bounded as a function of the risk of each of the analyses. We now explore what this could mean in terms of using differential privacy. Suppose we have a goal of maintaining differential privacy with $\epsilon = 0.1$. We begin performing analyses over the data, each with a smaller value of privacy parameter, say $\epsilon' = 0.01$. Performing the first analysis hence amounts to preserving differential privacy with privacy parameter 0.01. Performing two analyses amounts to preserving differential privacy with privacy parameter 0.02, and more generally performing k analyses amounts to preserving differential privacy with privacy parameter

$$k \cdot \epsilon' = k \cdot 0.01.$$

We get that performing $k = 10$ analyses amounts to preserving differential privacy with privacy parameter,

$$k \cdot \epsilon' = 10 \cdot 0.01 = 0.1 = \epsilon.$$

Performing further analyses would result in a privacy parameter that is larger (i.e., worse) than ϵ! It is sometimes useful to think about the use of differentially private computations as if each analysis results in spending a portion of an overall "privacy budget", possibly to the point where the "privacy budget" is exhausted and any further use would result in too high a privacy risk.

We note that in the above calculation we bounded the accumulated privacy risk by adding the privacy parameters of each analysis. It is possible to obtain better bounds on the accumulation of the privacy loss parameter, but this is beyond the scope of this discussion. A calculator for accumulated privacy risks is one of the tools that would be provided in the Privacy Tools project.

It is important to note that the fact privacy risk grows with use of data is not unique to differential privacy. In fact, this a fundamental law of information and hence applies to any disclosure control technique. The impression that statistical disclosure limitation techniques other than differential privacy do not suffer a similar accumulated degradation in privacy is merely due to the fact that these techniques are not generally analyzed with the same level of rigor that differential privacy is.

8 Tools for differentially private analysis

8.1 Differential privacy in Harvard's Dataverse project

Harvard's *Privacy Tools for Sharing Research Data* project develops tools to help enable the collection, analysis, and sharing of personal data for research in social science and other fields while

providing privacy protection for individual subjects. In particular, the project seeks to integrate the definitions and algorithmic tools from differential privacy into the Dataverse repository platform via tools such as TwoRavens[6] and DataTags[7].

Dataverse is a software application that enables institutions to host research data repositories. It provides a preservation and archival infrastructure for researchers to share data. One tool integrated with Dataverse is TwoRavens, a browser-based interface for exploring and analyzing data. The Privacy Tools project seeks to develop a differentially private version of the core functionality of TwoRavens. Through the differentially private access enabled by TwoRavens, researchers will be able to perform rough preliminary analyses of potentially sensitive datasets that currently cannot be safely shared. Such access will help researchers determine whether it is worth the effort to apply for full access to the raw data.

DataTags is a framework under development for integration with Dataverse. DataTags are simple, iconic labels that categorically describe data handling and sharing specifications. A DataTag can be assigned to a dataset based on the risk profile and handling policy associated with the data. Each DataTag maps to a different transfer, storage, and access level, from completely open personal data (a "green" tag) to highly restricted ("red" tag) with maximum control (a "crimson" tag). When a researcher seeks to deposit a dataset containing personal information into Dataverse, she will proceed through an automated interview that will assign a DataTag and construct a data use agreement to accompany the dataset. The Dataverse repository will guarantee that the storage and access requirements specified by the DataTag are met. The Privacy Tools project aims to use DataTags to denote the risk profiles and handling policies for different versions of a dataset or for statistics derived from a dataset. For example, while a raw version of a sensitive dataset might be assigned a restricted ("red" or "crimson") DataTag, differentially private statistics derived from the data might be assigned an open ("green") DataTag. Members of the project are also assessing the protection guaranteed by various settings of the differential privacy parameters (ϵ and δ) so that they can make recommendations regarding which parameters are appropriate for public or restricted access to datasets maintained in Dataverse.

8.2 Other implementations of differential privacy

Several experimental systems exist that allow analysts to construct analyses that preserve differentially private analyses while not concerning them with the complication of providing and managing differential privacy. These include Privacy Integrated Queries (PINQ)[8], Airavat[9] and GUPT.

9 Summary

Differential privacy provides us with a quantifiable measure of privacy. The quantification is via the privacy loss parameter – ϵ. The privacy loss parameter controls, simultaneously for every individual contributing to the analysis, the deviation between his or her privacy-ideal scenario and the real execution of the differentially private analysis. This deviation can grow as an individual participates in more analyses, but the overall deviation can be bounded as a function of ϵ and number of analyses.

[6] `http://datascience.iq.harvard.edu/about-tworavens`

[7] `http://datatags.org/`

[8] See `http://research.microsoft.com/en-us/projects/pinq/`.

[9] See `http://z.cs.utexas.edu/users/osa/airavat/`.

The parameter ϵ can be interpreted as bounding the excess risk to an individual resulting from her data being used (as compared with her risk when her data is not being used). Indirectly, ϵ also controls the accuracy to which the differentially private computation can be performed.

For researchers depositing sensitive data in Dataverse, the interface allowing them to choose from a variety of differentially private summary statistics while maintaining a desired level of privacy (in form of an accumulated privacy parameter) and then compute these summary statistics provides a tools for showcasing their data to other researchers that may be interested in using it in their studies, hence benefiting them with recognition within their community. For researchers seeking data for studies, the differentially private summary statistics would provide a means for deciding whether the data could be useful for them (and hence to proceed in a negotiation for obtaining the data) or not (saving them the negotiation and wait).

10 Advanced Topics

We conclude with a few more advanced topics that may be of interest.

10.1 Composition attacks – why do small leakages matter?

Consider the following hypothetical example.

> Alice and Bob are both professors at Harvard. They have both been given access to the student database, which contains sensitive information, including financial aid data. Gaining access required much effort on their part, such as undergoing special training and signing strict privacy agreements.
>
> Alice publishes a paper in March in which she reveals that "currently Harvard has 1000 freshmen and their parents on average earn $\$100,000$ per year." Alice reasons that, since the number she revealed is an average taken over 1000 people, no individual's private information is exposed. In April, Bob publishes another paper in which he states that "the average parental income of Harvard's 999 freshmen is $\$99,000$ per year." Alice and Bob are not aware that both of them have revealed similar information.
>
> A clever student Eve reads both Alice and Bob's papers and notices the discrepancy. From the revealed information, Eve concludes that between March and April one freshman left Harvard and that student's parents earn $1000 \times \$100,000 - 999 \times \$99,000 = \$1,099,000$ per year. Eve asks around and is able to determine that a student named John dropped out around the end of March. Eve then informs others that she determined that John's parents earn $\$1,099,000$ per year.
>
> John is understandably upset when it is revealed exactly how wealthy his family is. He complains to the university and Alice and Bob are asked to explain. Both Alice and Bob argue that the information they revealed did not identify any individuals. (Bob argues that his statement satisfied 999-anonymity and therefore protected the privacy of the students concerned.) Alice and Bob were totally unaware of each other's statements.

This story illustrates the danger posed by *composition* – that is, combining the results of multiple supposedly-private analyses on the same individuals to draw conclusions. Both Alice and Bob revealed information that, in isolation, seems innocuous. However, in conjunction, their statements severely compromised John's privacy.

Such a privacy breach is difficult for Alice or Bob to avoid individually, as they do not know what information has already been revealed or will be revealed by others in future.

However, if Alice and Bob had both revealed differentially private statistics, then this situation could have been avoided: Suppose Alice distorted the average income she revealed in accordance with satisfying differential privacy with privacy parameter $\epsilon = 0.1$. Likewise, suppose Bob also distorted the average income he revealed so as to satisfy differential privacy with privacy parameter $\epsilon = 0.1$. Then – even though Alice and Bob did not collaborate to ensure privacy – we can be sure that combining the information they revealed still satisfies differential privacy with privacy parameter $\epsilon = 0.2$.

This example illustrates one of the greatest strengths of differential privacy: If multiple analyses are conducted on the same set of individuals, then, as long as each analysis *individually* satisfies differential privacy, we can be assured that all of the information released, when viewed together, will remain differentially private.

However, each analysis incurs some privacy risk to the individuals concerned and this risk accumulates. Thus there is a limit to how many analyses can be performed. This is why it is important to quantify risk to privacy and to understand quantitatively how risk can accumulate as we have shown in sections 3.1 and 7.2 above.

We note that while differential privacy is not the only technique that quantifies risk, it is currently the only framework with quantifiable guarantees on how risk accumulates from a composition of several private analyses.

10.2 Group privacy

By referring to individuals' privacy-ideal scenarios, the definition of differential privacy directly addresses information that is localized to an individual. We can proceed in this manner to also consider privacy-ideal scenarios for a couple of individuals, for example we can consider the scenario where both John's and Gertrude's information is omitted from the input to the analysis - this would be John and Gertrude's privacy-ideal scenario. It is particularly relevant to consider this scenario if John and Gertrude are related, as they may share many attributes (such as their address) and so John's privacy may be connected to Gertrude's.

Similarly, we can consider the group-privacy-ideal scenario for a group of k individuals - it would simply be the scenario where the information of all k individuals is omitted from the input to the analysis.

Recall that the parameter ϵ controls how "distant" the real scenario can be from any individual's privacy-ideal scenario. It can be shown that for the group-privacy-ideal setting of a group of k individuals, this distance grows to at most

$$k \cdot \epsilon.$$

This means that the privacy guarantee degrades moderately with the size of the group. Effectively, a meaningful privacy guarantee is provided to groups of individuals of size up to about

$$k \approx 1/\epsilon$$

individuals. However, almost no protection is guaranteed to information that is not localized to less than

$$k \approx 10/\epsilon$$

individuals. This reflects a design choice in differential privacy not to a-priori exclude learning trends in moderately-sized groups.

10.3 Amplifying privacy – secrecy of the sample

Given an analysis A that preserves differential privacy with a privacy loss parameter ϵ it is possible to create an analysis A' that has a better loss parameter $\epsilon' < \epsilon$.

Suppose your data is a random and secret sample from a larger population, where secret means that the choice of the people in the sample has not been revealed. Then the accuracy of your differentially private analysis without changing the privacy guarantee. Estimate the size of the larger population. It is important to be conservative in your estimate. In other words, it is okay underestimate but could violate privacy if you overestimate.

10.4 Differential privacy – a property of the analysis (not its specific outcome)

Many disclosure limitation techniques restrict the outcome of a computation rather the computation itself. For example, the data anonymization technique known as k-anonymity transformed data in tabular form so as the information for each person contained in the k-anonymized data is identical to that of at least $k-1$ other individuals whose information appears in the k-anonymized table. In other words, k-anonymity is a property of the anonymized data, and no further restrictions are put on the process of creating a k-anonymized data. Note that in general many k-anonymized tables exist for a given dataset. A hypothetical k-anonymizer may (maliciously or inadvertently) choose among these tables in a way that depends on an individual's sensitive attribute. For example, if for the given data there exist two possible output tables T_1 and T_2, the k-anonymizer may choose to output T_1 if John is HIV+ and T_2 otherwise, hence compromising John's privacy! While we do not claim real implementations of k-anonymity suffer from this problem, the notion of k-anonymity does not preclude it.

The observant reader may have noticed that the requirement of differential privacy is of a very different flavor. Instead of restricting the outcome of a differentially private computation, the definition restricts the process producing it. We further delve on this issue to make it more intuitive.

Consider what happens when a statistical analysis is performed over sensitive data, as in Figure 1. To be of any interest, the outcome of the analysis must depend on the data and hence the computation of the analysis must exhibit some non-zero leakage of information, which means there must be some risk to privacy. The concern for privacy is that an individual or organisation observing the outcome of this computation would use it to infer sensitive individual data. It is important to understand why such breach of privacy can happen. We consider a few illustrative examples.

Suppose that a computation performed on medical records including John's outputs the line: John, HIV+. Does that mean that John's privacy was breached as a result of this computation (in the sense that some sensitive information about John was revealed)? The answer is that this is not necessarily the case! For example, if the computation ignores its input data and always output John, HIV+. In this case, there is no functional dependence between the HIV status in John's medical records and the outcome, and hence this mechanism does not leak any information about John (although John may have other reasons for objecting to its deployment).

As another extreme example, omitting John from the outcome is not by itself a guarantee for his privacy. To see how a privacy breach may be caused, consider a mechanism that redacts all

records of HIV+ patients. An observer that knows John is a patient and notices that his record is not included in the outcome can learn that John is HIV+.

More generally, the outcome of a mechanism applied to sensitive data can depend on its input data in much more complicated ways than those described above, and the relationship between input data and outcome can even be randomized. But the intuition underlying our examples still holds: it is the functional relationship between input and output that the computation implements which determines to what extent sensitive information can be learned from the outcome.

The definition of differential privacy follows this intuition closely. Differential privacy is not a property of a specific outcome, but a property that an analysis (or computation) may or may not have. To satisfy differential privacy, the analysis' behavior should not change noticeably in case John's (or any other single individual's) data is added to or removed from the input database.

10.5 Differential privacy vs. opt-out mechanisms

11 Further reading

Differential privacy was introduced in 2006 by Dwork, McSherry, Nissim and Smith [3]. Our presentation of privacy-ideal scenario vs. real computation is influenced by [1] and our risk analysis is influenced by [6]. For other presentations of differential privacy see [2] and [5]. For a thorough technical introduction into differential privacy see [4].

References

[1] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[2] Cynthia Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54(1):86–95, 2011.

[3] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In 3^{rd} *Theory of Crypt. Conf.*, pages 265–284, 2006.

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

[5] Ori Heffetz and Katrina Ligett. Privacy and data-based research. *Journal of Economic Perspectives*, 28(2):75–98, 2014.

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