WP. 26
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

Differential Privacy – A Primer for the Perplexed

Prepared by Cynthia Dwork and Frank McSherry (Microsoft Research, U.S.A.), Kobbi Nissim (Ben-Gurion University, Israel) and Adam Smith (Pennsylvania State University, U.S.A.)

Differential Privacy: A Primer for the Perplexed

Cynthia Dwork*, Frank McSherry*, Kobbi Nissim** and Adam Smith***

* Microsoft Research Silicon Valley, USA.

** Ben-Gurion University, Israel.

*** Pennsylvania State University, USA.

Abstract. Differential privacy is a definition of a privacy goal tailored to privacy-preserving data analysis. Since its inception [DMNS06, Dwo06], differentially private data analysis has been the subject of intense algorithmic and theoretical investigation. Unfortunately, some of the literature, however, suffers from basic misconceptions about the definition. In this brief invited note we elucidate and respond to the most common of these errors.

1 Error 1: Confusing a definition with a particular algorithm

Differential privacy is a definition. It is a mathematical guarantee that can be satisfied by an algorithm that releases statistical information about a data set. Many different algorithms satisfy the definition.

In the sequel we will assume there is a database X containing n rows, each with data belonging to a single individual. If each individual's data is described as an element of a set U, then we may think of a data set as a multi-set in U (that is, a set in which each element may appear more than once). Let $\mathbb{D}$ denote the space of possible data sets. There is an algorithm M that takes as inputs the database and, optionally, a query, and returns a response. We refer to such algorithms as *mechanisms.* We will state the definition of differential privacy presently. For now, however, we note that if a randomized mechanism M has the same output distribution on all databases, then its output reveals no information about its input. In fact, this is equivalent to the following statement, where $\mathbb{D}$ is the set of all possible databases and Q is the set of all possible queries:

$$\forall X, X' \in \mathbb{D}, \forall q \in Q, \forall S \subseteq \text{Range}(M) : \frac{\Pr[M(X,q) \in S]}{\Pr[M(X',q) \in S]} = 1 = e^0, \tag{1}$$

where in both the numerator and the denominator the probability space is *over the random choices made by the mechanism.* Note that a deterministic mechanism can also satisfy this property, if it ignores its input.

It follows from Equation 1 that if an output y is produced, this reveals no information about the database. Equation 1 is absolute: If an algorithm has this property, then it holds without regard to what the observer – the data analyst who sees the output y – knows or has access to. If the observer already knows that the database is X, or that the database contains the medical data of the authors of this note, or even if the observer knows the modulo 2 sum of all sickle cell status bits of the people in the database, this cannot change the *behavioral* guarantee of the mechanism. Interacting with a mechanism that satisfies Equation 1 *cannot* leak any further information about the database.

This is a definition of perfect secrecy, advanced in the context of encryption by Shannon [Sha49]. It is not an algorithm. Rather, it is a condition on the probability distribution of randomized algorithms; any algorithm that satisfies this condition reveals no information, in an information-theoretic sense, about the data.

The symmetric difference between two databases X and X', denoted $X \ominus X'$, is the set of elements, or rows, that appear in either X or X' but not in their intersection. For example, if X' is a subset of X and X contains just one row not present in X', then the symmetric difference has cardinality 1. (Since X and X' can be multi-sets, the symmetric difference is actually a bit more complicated: an item appears k times in $X \ominus X'$ if it appears k more times in one database than the other). We are now ready to state the definition of differential privacy. It has exactly the same form as Equation 1, only now we permit a little bit of information to be released:

Definition 1. *A mechanism satisfies* ε-differential privacy *if for every pair of databases* X, X'*, and for every subset* $S \subseteq$ *Range*(M)*, and every query* $q \in Q$ *(if the optional "query" argument is used),*

$$\frac{\Pr[M(X, q) \in S]}{\Pr[M(X', q) \in S]} = e^{\varepsilon |X \ominus X'|} \tag{2}$$

where in both the numerator and the denominator the probability space is over the random choices made by the mechanism.

This definition differs from the definition of perfect secrecy (Equation 1) in exactly one way: the exponent of e in the bound on the probability ratio is $\varepsilon |X \ominus X'|$ rather than 0. If an algorithm satisfies Definition 1, then this holds without regard to what the observer – the data analyst who sees an output y – knows or has access to. If the observer already knows that the database is X, or that the database contains the medical data of the authors of this note, or even if the observer knows the modulo 2 sum of all sickle cell status bits of the people in the database, this cannot change the *behavioral* guarantee of the mechanism.

One common interpretation of this definition is that no matter what an observer (or attacker) knows ahead of time, the output of a differentially private algorithm gives little indication of whether any particular given user's data was included in the data set. We return to this point later.

For now, we summarize: the definition is not an algorithm, but rather a condition satisfied by many different algorithms. Note that formulating privacy in these terms, as a requirement that can be satisfied in several ways, provides a framework where one can study algorithms, compare their privacy guarantees, and understand their joint effect on privacy. We believe it is a necessary step in a scientific approach to privacy (see, e.g., [KL10] for a discussion of other definitions of this form).

2 Error 2: Confusing outputs with the process that generates them

Consider the famous *1-bit one-time pad* mechanism, which yields perfect secrecy. The input to this randomized mechanism is a bit $b \in \{0, 1\}$. The mechanism chooses a new random bit $p \in_R \{0, 1\}$ and outputs $b + p \bmod 2$.

Claim 1. *The 1-bit one-time pad mechanism yields perfect secrecy.*

Proof. The proof is immediate because the mechanism satisfies Equation 1. Spelling it out: the number of people n is 1 and the space of databases is $\mathbb{D} = \{0, 1\}$; there is no "query" argument, and the range of the mechanism is $\{0, 1\}$. By exhaustively checking all the possibilities, we see that $1/2 = \Pr[M(0) = 1] = \Pr[M(0) = 0] = \Pr[M(1) = 1] = \Pr[M(1) = 0]$. Thus,

$$\forall X, X' \in \mathbb{D}, \forall q \in Q, \forall S \subseteq \text{Range}(M) : \frac{\Pr[M(X, q) \in S]}{\Pr[M(X', q) \in S]} = 1. \quad \square$$

Notice that although the mechanism actually computed using its input, so in some formal operation sense it does not ignore its input, the effect is the same: the output distribution is the same, independent of the database.

As noted above, this means that the output of the mechanism yields no information about the database. This is true if the output is 0 and the database is 0, the output is 1 and the database is 0, the output is 0 and the database is 1, or the output is 1 and the database is 1. In all four cases there is nothing to be learned from the output about the database. It is the process by which the outputs are generated that ensures this. Thus, we do not speak of a "safe" or "privacy-preserving" output; rather, we think of outputs that have been generated according to a privacy-preserving process. The 1-bit one-time pad is a process that yields perfect privacy.

All perfectly private mechanisms are differentially private, but the class of differentially private mechanisms contains mechanisms that do not yield perfect privacy; rather, they provide some nontrivial information about the database, as we would hope if we are interested in privacy-preserving data analysis.

Randomized response is such a mechanism [War65]. Consider the following *1-bit randomized response* mechanism. As with the 1-bit one-time pad mechanism, the set of possible databases is $\mathbb{D} = \{0, 1\}$. Let $X \in \{0, 1\}$ be the actual database. As in the 1-bit one-time pad mechanism, a bit $p \in_R \{0, 1\}$ is chosen. Now, however the mechanism proceeds as follows. If $p = 1$, then a second coin $c \in_R \{0, 1\}$ is flipped and the result is published. If $p = 0$ then release X is published.

Claim 2. *The 1-bit randomized response mechanism is* $(\ln 3)/2$*-differentially private.*

Proof. Fix $d \in \{0, 1\}$. Without loss of generality, let $X = d$ and $X' = 1 - d$. On input X, d is output with probability $3/4$. On the other hand, on input X', d is output with probability $1/4$. The ratio is 3. The symmetric difference difference is 2, yielding $((\ln 3)/2)$-differential privacy. $\square$

Note that there is no such thing as a "good" output or a "bad output" for a given database. Nonetheless, the released value does give a little bit of a hint about which of the two 1-bit databases is more likely, but the randomness provides uncertainty. And as usual we point out that the 1-bit randomized response mechanism retains its differential privacy property independent of what anyone seeing the output may know about the database or its connection to other databases.

3 Error 3: Confusing poor utility of a specific differentially private algorithm with a failure of differential privacy *per se*

As we have noted, the 1-bit perfect one-time pad is a differentially private algorithm. More specifically, for all $\varepsilon \geq 0$ the 1-bit one-time pad algorithm is ε-differentially private. It also yields *no* information about the database. This does not mean that "differential privacy can't work," or that no differentially private algorithm can produce useful outputs. Indeed, we have also seen that 1-bit randomized response is $(\ln 3)/2$-differentially private. From the point of view of utility it is a *better* $(\ln 3)/2$-differentially private algorithm than the one-time pad. In particular, it is well known that, assuming all n respondents in a randomized-response survey follow the instructions correctly, one can derive from their answers an estimate for the number of respondents admitting to a particular behavior with expected *inaccuracy* (or distortion) on the order of $\sqrt{n}$.

In fact, there are $(\ln 3)/2$-differentially private algorithms with even better accuracy, namely, yielding an estimated count with expected distortion independent of n. When $\varepsilon = (\ln 3)/2$, the *Laplace mechanism* [DMNS06] releases a tally with the correct mean and standard deviation $\sqrt{2}(\sqrt{3}/2) = \sqrt{3/2}$, instead of standard deviation that is $\Theta(\sqrt{n})$ as in randomized response.

Thus, the limitations of a particular differentially private algorithm don't necessarily apply to all differentially private algorithms.[1] Any claim that differential privacy cannot achieve a given accuracy goal *must* be accompanied by a proof showing that differential privacy – the specific requirement on ratios of probability distributions – is inconsistent with the goal. That is, one must show that *no* differentially private algorithm can possibly achieve a given level of accuracy. This is quite tricky and our intuition about what is or is not possible is often wrong. For example, it is natural to conjecture that the Laplace mechanism cited above is basically optimal for answering "tally" type questions differentially privately. However, in settings where many tallies must be computed simultaneously, there are more complicated algorithms (e.g., [BLR08]) that do much better. Despite these difficulties, there are a few published results that prove non-trivial limitations of differentially private algorithms (see [GRS08, HT09] for particularly elegant examples).

4 Miscellaneous "Does-es" and "Does Nots" of Differential Privacy

We reiterate our main points thus far:

1. *Differential privacy is a definition, not an algorithm.*
2. *For any specific database, no outputs are "good" or "bad". Privacy (perfect or differential) is obtained from the process by which the outputs are generated.*
3. *Failure of a specific ε-differentially private mechanism to provide sufficiently accurate answers does not imply that differential privacy is incompatible with this goal.*

In this last section we elaborate a bit on these points and further explicate differential privacy, addressing additional misconceptions seen in the literature regarding several of

[1] Similarly, problems with a particular *implementation* of a given algorithm do not necessarily imply that the algorithm itself is incorrect.

the things that differential privacy does or does not provide.

Differential privacy is commonly interpreted in the context of sensitive records supplied to a data analysis. The privacy guarantee is that the distribution over outputs, and consequently the distribution over conclusions drawn from those outputs, varies by at most an (often small) multiplicative factor when an individual opts in to or opts out of the input database. In consequence is that there is little harm in participating; any conclusions that will be reached about you, your data, or the world at large will be almost as likely to be reached without your data as with it.

Fact: Differential privacy does not guarantee that what you believe are your secrets remain secret. What differential privacy guarantees is that your participation in a survey will not disclose specifics that you contributed to the survey. It is very possible that conclusions drawn from the survey may reflect statistical information about you. A health survey intended to discover early indicators of a particular ailment may produce strong, even conclusive results; that these conclusions hold for you is not evidence of a differential privacy violation. Indeed, you may not even have participated in the survey (again, differential privacy ensures that these conclusive results would be obtained with very similar probability whether or not you participated in the survey). In particular, if the survey teaches us that specific private attributes correlate strongly with public attributes, this is not a violation of differential privacy, since this same correlation would be observed with almost the same probability independent of the presence or absence of any respondent. Our favorite parable for this point is:

> Suppose a statistical database teaches us that smoking causes cancer. Alice, a known smoker, is harmed because her insurance premiums rise. Alice would be (essentially) just as likely to be harmed independent of whether or not she is in the database. Alice is also helped: because of the study she enters a smoking cessation program.

Differential privacy was designed to ensure that Alice and everyone else is motivated to join these socially useful studies.

Error: Differential privacy "does not make sense" for $\epsilon > 1$. It is (oddly) often claimed that, as probabilities are at most one in value, values of ε greater than one make little sense. It is unclear where this specific choice comes from, as $\exp(1) = \mathrm{e}$ has no particular relationship to the maximum value of probabilities, although the spirit that large epsilon are fundamentally meaningless is understandable. Certainly, larger values of epsilon result in substantially less meaningful guarantees than smaller values of epsilon. However, they still give mathematically meaningful guarantees. Consider the release of search logs by AOL, in which an individual was identified by her searches. What would the probability of the event that a New York Times reporter would publish an article accurately describing the private worries and concerns of Thelma Arnold, had her data not been included in the data set? Suppose we guess one out of a trillion. In such a case a guarantee of the form "Allowing your data to be included increases your risk of embarrassment by a factor of $e^{12} < 164,000$" still leaves disclosure very unlikely.

Error: Differential privacy cannot work because no one can explain how to set the value of ε. It is a strength, not a weakness, of differentially private methods that it is possible to maintain a quantitative measure of the cumulative privacy loss suffered by an individual in a *given* database. It is a further strength, not a weakness, of the research in differential privacy that we have a good understanding of the cumulative privacy loss suffered by an individual whose data are in *multiple, independently operated* databases, even against an adversary that has access to all these databases. It is an additional further strength, not a weakness, that differential privacy research has developed technology to dramatically reduce privacy loss when coordinated answers are possible. It is yet one more additional further strength, not a weakness, that differential privacy researchers have established that such coordination is essential for enabling very high complexity queries, and cannot be replaced. It is this line of investigation – in differential privacy and in no other approach to private data analysis – that allows us, through ε, to understand and to mitigate privacy loss. Yes, this research is incomplete. Yes, theorems of the following form seem frighteningly restrictive:

> If an individual participates in $10,000$ adversarially chosen databases, and if we wish to ensure that her cumulative privacy loss will, with probability at least $1 - e^{-32}$, be bounded by e^1, then it is sufficient that each of these databases will be $\varepsilon = 1/801$-differentially private.

But how else can we find a starting point for understanding how to relax our worst-case adversary protection? How else can we measure the effect of doing so? And what other technology permits one to prove such a claim?

On a more philosophical level, consider an analogy to time: there are only so many hours in your lifetime, and once they are consumed you die. (This is sometimes worse than someone learning private information about you.) Yet, somehow, we as a society have found ways to arrive at values for an individual's time, and a fundamental part of that is the ability to quantitatively measure it.

There is certainly new thinking to do about the value of privacy, but we believe it is an exceptionally good thing that one can now do this; every meaningful privacy guarantee should be similarly quantitative.

References

[BLR08] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th ACM SIGACT Symposium on Thoery of Computing*, 2008.

[DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284, 2006.

[Dwo06] C. Dwork. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)(2)*, pages 1–12, 2006.

[GRS08] A. Ghosh, T. Roughgarden, and M. Sundarajan. Universally utility-maximizing privacy mechanisms. Manuscript, November 2008.

[HT09] M. Hardt and K. Talwar. On the geometry of differential privacy. arXiv:0907.3754v2, 2009.

[KL10] Daniel Kifer and Bing-Rong Lin. Towards an axiomatization of statistical privacy and utility. In Jan Paredaens and Dirk Van Gucht, editors, *PODS*, pages 147–158. ACM, 2010.

[Sha49] Claude Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656–715, 1949.

[War65] S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *JASA*, pages 63–69, 1965.