Vanderbilt Journal of Entertainment & Technology Law

Volume 16
Issue 4 *Issue 4 - Summer 2014*

Article 1

2014

Fool's Gold: An Illustrated Critique of Differential Privacy

Jane Bambauer

Krishnamurty Muralidhar

Rathindra Sarathy

VANDERBILT JOURNAL OF
ENTERTAINMENT & TECHNOLOGY LAW

VOLUME 16 SUMMER 2014 NUMBER 4

Fool's Gold: An Illustrated Critique of Differential Privacy

Jane Bambauer, Krishnamurty Muralidhar,** and Rathindra Sarathy****

ABSTRACT

Differential privacy has taken the privacy community by storm. Computer scientists developed this technique to allow researchers to submit queries to databases without being able to glean sensitive information about the individuals described in the data. Legal scholars champion differential privacy as a practical solution to the competing interests in research and confidentiality, and policymakers are poised to adopt it as the gold standard for data privacy. It would be a disastrous mistake.

This Article provides an illustrated guide to the virtues and pitfalls of differential privacy. While the technique is suitable for a narrow set of research uses, the great majority of analyses would produce results that are beyond absurd—average income in the negative millions or correlations well above 1.0, for example.

The legal community mistakenly believes that differential privacy can offer the benefits of data research without sacrificing privacy. In fact, differential privacy will usually produce either very wrong research results or very useless privacy protections. Policymakers and data stewards will have to rely on a mix of

* Associate Professor of Law, University of Arizona, James E. Rogers College of Law; J.D., Yale Law School; B.S., Yale College.

** Gatton Research Professor, University of Kentucky, Gatton College of Business & Economics; Ph.D., Texas A&M University; M.B.A., Sam Houston State University; B.Sc. University of Madras, India.

*** Ardmore Chair, Oklahoma State University, Spears School of Business; Ph.D., Texas A&M University; B.E., University of Madras, India.

approaches—perhaps differential privacy where it is well suited to the task and other disclosure prevention techniques in the great majority of situations where it isn't.

TABLE OF CONTENTS

INTRODUCTION

A young internist at the largest hospital in a midsized New England city is fretting. She has just diagnosed an emergency room patient with Eastern Equine Encephalitis Virus (EEEV). The diagnosis troubles the internist for a number of reasons. Modern medicine offers neither a vaccine nor an effective treatment.[1] Moreover, the internist remembers that a colleague diagnosed a different patient with EEEV three weeks ago and knows that there was a third case a few weeks before that. The disease is transmitted by mosquitos and is not communicable between humans. However, an influx of cases would suggest that the local mosquito population has changed, putting the city's inhabitants at risk. So, the internist is fretting about whether the three cases that have come through the hospital in the last six weeks merit a phone call to the state and national centers for disease control.

1. *See Eastern Equine Encephalitis*, Centers for Disease Control & Prevention, http://www.cdc.gov/EasternEquineEncephalitis/index.html (last updated Aug. 16, 2010).

To aid her decision, the internist decides to query a state health database to see how many cases of the rare disease have occurred in her city in each of the last eight years. Recently, the state health database proudly adopted differential privacy as a means to ensure confidentiality for each of the patients in the state's database.

Differential privacy is regarded as the gold standard for data privacy.[2] To protect the data subjects' sensitive information, differential privacy systematically adds a random number generated from a special distribution centered at zero to the results of all data queries. The "noise"—the random value that is added—ensures that no single person's inclusion or exclusion from the database can significantly affect the results of queries. That way, a user of the system cannot infer anything about any particular patient. Because the state health department is also concerned about the utility of the research performed on the database, it has chosen the lowest level of noise recommended by the founders of differential privacy. That is to say, the state has chosen the *least* privacy-protecting standard in order to preserve as much utility of the dataset as possible.

When the internist submits her query, the database produces the following output:[3]

Query = Count of Patients Diagnosed with EEEV within the City

Year	N		Year	N
2012	837.3		2007	5,019.3
2011	211.3		2006	868.6
2010	−794.6		2005	−2,820.6
2009	−1,587.8		2004	2,913.9
2008	2,165.5			

What is the internist to make of this data?

2. *See* Raghav Bhaskar et al., *Noiseless Database Privacy*, *in* ADVANCES IN CRYPTOLOGY – ASIACRYPT 2011: 17TH INTERNATIONAL CONFERENCE ON THE THEORY AND APPLICATION OF CRYPTOLOGY AND INFORMATION SECURITY 215, 215 (Dong Hoon Lee & Xiaoyun Wang eds., 2011); Samuel Greengard, *Privacy Matters*, 51 COMMC'NS OF THE ACM, Sept. 2008, at 17, 18; Graham Cormode, *Individual Privacy vs Population Privacy: Learning to Attack Anonymization*, *in* KDD'11 Proceedings of the 17th ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1253, 1253 (2011). *But see* Fida K. Dankar & Khaled El Emam, *Practicing Differential Privacy in Health Care: A Review*, 6 TRANSACTIONS ON DATA PRIVACY 35, 51–60 (2013) (noting theoretical limitations that differential privacy must address before it can be widely adopted for health care research).

3. This is an actual instantiation of the differential privacy technique. The noise in this exercise was randomly drawn after setting $\varepsilon = ln(3)$ and allowing for 1,000 queries to the database. For a description of the technique, see *infra* Part I.B.

If the internist is unfamiliar with the theory behind differential privacy, she would be baffled by the respones. She would be especially puzzled by the negative and fractional values since people do not tend to be negative or partial.[4] The internist is likely to conclude the responses are useless, or worse, that the system is seriously flawed.

If the internist happens to be familiar with the theory behind differential privacy, she would know that there is a very good chance—to be precise, a 37% chance—that the system is adding over 1,000 points of noise in one direction or the other. However, even knowing the distribution of noise that is randomly added to each cell, the internist has no hope of interpreting the response. The true values could be almost anything. It could be that the city has consistently diagnosed dozens of patients a year with EEEV, rendering her experience little reason for alarm. Or it could be that the true values are all zero, suggesting that there is reason for concern. The noise so badly dwarfs the true figures that the database query is a pointless exercise.

This hypothetical is a representative example of the chaos that differential privacy would bring to most research database systems. And yet, differential privacy is consistently held up as the best solution to manage the competing interests in privacy and research.[5]

Differential privacy has been rocking the computer science world for over ten years and is fast becoming a crossover hit among privacy scholars and policymakers.[6] Lay descriptions of differential privacy are universally positive. *Scientific American* promises that "a mathematical technique called 'differential privacy' gives researchers access to vast repositories of personal data while meeting a high standard for privacy protection."[7] Another journal, *Communications of the ACM*, describes differential privacy in slightly more detailed and equally appealing terms:

> Differential privacy, which first emerged in 2006 (though its roots go back to 2001), could provide the tipping point for real change. By introducing random noise and ensuring that a database behaves the same—independent of whether any individual or

4. *See* MICROSOFT, DIFFERENTIAL PRIVACY FOR EVERYONE 4–5 (2012), *available at* http://www.microsoft.com/en-us/download/details.aspx?id=35409 ("Thus, instead of reporting one case for Smallville, the [query system] may report any number close to one. It could be zero, or ½ (yes, this would be a valid noisy response when using DP), or even −1.").

5. *See* Bhaskar et al., *supra* note 2, at 215; Cormode, *supra* note 2, at 1253–54; Greengard, *supra* note 2, at 18.

6. Google Scholar has indexed over 2,500 articles on the topic. Google Scholar, www.scholar.google.com (last visited Apr. 12, 2014) (describing a search for "Differential Privacy").

7. Erica Klarreich, *Privacy By the Numbers: A New Approach to Safeguarding Data*, SCI. AM. (Dec. 31, 2012), http://www.scientificamerican.com/article/privacy-by-the-numbers-a-new-approach-to-safeguarding-data.

> small group is included or excluded from the data set, thus making it impossible to tell which data set was used—it's possible to prevent personal data from being compromised or misused.[8]

Legal scholars have also trumpeted the promise of differential privacy. Felix Wu recommends differential privacy for some scientific research contexts because the query results are "unreliable with respect to any one individual" while still making it sufficiently reliable for aggregate purposes.[9] Paul Ohm explains differential privacy as a process that takes the true answer to a query and "introduces a carefully calculated amount of random noise to the answer, ensuring mathematically that even the most sophisticated reidentifier will not be able to use the answer to unearth information about the people in the database."[10] And Andrew Chin and Anne Klinefelter recommend differential privacy as a best practice or, in some cases, a legal mandate to avoid the reidentification risks associated with the release of microdata.[11]

Policymakers have listened. Ed Felten, the chief technologiest for the Federal Trade Commission, praises differential privacy as "a workable, formal definition of privacy-preserving data access."[12] The developers of differential privacy have even recommended using the technique to create privacy "currency," so that a person can understand and control the extent to which their personal information is exposed.[13]

These popular impressions give differential privacy an infectious allure. Who *wouldn't* want to maximize database utility while ensuring privacy?

The truth, of course, is that there is no simple solution to the eternal contest between data privacy and data utility. As we will show, differential privacy in its pure form is a useful tool in certain

---

8. Greengard, *supra* note 2, at 18.

9. Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1139–40 (2013).

10. Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1756 (2010). Ohm acknowledges that differential privacy techniques add significant administration costs, and also risks denying the researcher an opportunity to mine the raw data freely to find useful patterns. *Id.* These are external critiques. Ohm does not present the internal critique of differential privacy theory that we develop here. *See id.*

11. Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*, 90 N.C. L. REV. 1417, 1452–54 (2012).

12. Ed Felten, *What Does it Mean to Preserve Privacy?*, TECH@FTC (May 15, 2012, 4:47 PM), http://techatftc.wordpress.com/2012/05/15/what-does-it-mean-to-preserve-privacy.

13. *See* Frank D. McSherry, *Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis*, *in* SIGMOD'09: PROCEEDINGS OF THE 2009 ACM SIGMOD International Conference on Management of Data 19, 25 (2009); Klarreich, *supra* note 7.

narrow circumstances. Unfortunately, most research occurrs outside of those circumstances, rendering a pure form of differential privacy useless for most research. To make differential privacy practical for the vast majority of data research, one would have to diverge significantly from differential privacy's pure form.

Not surprisingly, this is the direction in which advocates of differential privacy have gone.[14] It is the only way *to* go if one harbors hopes for general application of the technique. But the only way to convert differential privacy into a useful tool is to accept and adopt a range of compromises that surrender the claim of absolute "ensured" privacy. In other words, a useful version of differential privacy is not differential privacy at all. It is a set of noise-adding practices indistinguishable in spirit from other disclosure prevention techniques that existed well before differential privacy burst onto the scene. Thus, differential privacy is either not practicable or not novel.

This Article provides a comprehensive, but digestible, description of differential privacy and a study and critique of its application. Part I explains the age-old tension between data confidentiality and utility and shows how differential privacy strives to thread the needle with an elegant solution. To this end, Part I recounts a brief history of the development of differential privacy and presents a successful application of differential privacy that demonstrates its promise.

Part II explores the many contexts in which differential privacy *cannot* provide meaningful protection for privacy without sabotaging the utility of the data. Some of the examples in this section are lifted directly from the differential privacy literature, suggesting, at least in some cases, that the proponents of differential privacy do not themselves fully understand the theory. The most striking failures of differential privacy (correlations greater than 1, average incomes in the negative millions) track some of the most general, common uses of data. Part II demonstrates clearly that differential privacy cannot serve as the lodestar for the future of data privacy.

Part III conducts a postmortem. What went wrong in the applications of differential privacy described in Part II? Looking forward, how can we know in advance whether differential privacy is a viable tool for a particular research problem? The answers provide insight into the limitations of differential privacy's theoretical underpinnings. These limitations can point researchers in the right direction, allowing them to understand when and why a deviation

14. *See* Bhaskar et al., *supra* note 2, at 215–16; Cynthia Dwork & Adam Smith, *Differential Privacy for Statistics: What We Know and What We Want to Learn*, 1 J. PRIVACY & CONFIDENTIALITY 135, 139 (2009).

from the strict requirements of differential privacy is warranted and necessary. We also identify and correct some misinformed legal scholarship and media discussion that give unjustified praise to differential privacy as a panacea.

The Article concludes with a dilemma. On one hand, we praise some recent efforts to take what is good about differential privacy and modify what is unworkable until a more nuanced and messy—but ulitimately more useful—system of privacy practices are produced. On the other hand, after we deviate in important respects from the edicts of differential privacy, we end up with the same disclosure risk principles that the founders of differential privacy had insisted needed to be scrapped. In the end, differential privacy is a revolution that brought us more or less where we started.

I. WHAT IS DIFFERENTIAL PRIVACY?

Protecting privacy in a research database is tricky business. Disclosure risk experts want to preserve many of the relationships among the data and make them accessible.[15] This is a necessary condition if we expect researchers to glean new insights. However, the experts also want to thwart certain types of data revelations so that a researcher who goes rogue—or who was never really a researcher to begin with—will not be able to learn new details about the individuals described in the dataset. How to preserve the "good" revelations while discarding the "bad" ones is a puzzle that has consumed the attention of statisticians and computer scientists for decades.[16]

When research data sets are made broadly available for research purposes, they usually take one of two forms.[17] Sometimes

15. *See* George T. Duncan & Sumitra Mukherjee, *Optimal Disclosure Limitation Strategy in Statistical Databases: Deterring Tracker Attacks through Additive Noise*, 95 J. OF THE AM. STAT. ASS'N 720, 720 (2000); Krishnamurty Muralidhar et al., *A General Additive Data Perturbation Method for Database Security*, 45 MGMT. SCI. 1399, 1399–1401 (1999); Krishnamurty Muralidhar & Rathindra Sarathy, *Data Shuffling—A New Masking Approach for Numerical Data*, 52 MGMT. SCI. 658, 658–59 (2006) [hereinafter Muralidhar & Sarathy, *Data Shuffling*]; Rathindra Sarathy et al., *Perturbing Nonnormal Confidential Attributes: The Copula Approach*, 48 MGMT. SCI. 1613, 1613–14 (2002); Mario Trottini et al., *Maintaining Tail Dependence in Data Shuffling Using t Copula*, 81 STAT. & PROBABILITY LETTERS 420, 420 (2011).

16. "Statistical offices carefully scrutinize their publications to insure that there is no disclosure, i.e., disclosure of information about individual respondents. This task has never been easy or straightforward." I. P. Fellegi, *On the Question of Statistical Confidentiality*, 67 J. AM. STAT. ASS'N 7, 7 (1972).

17. These two popular forms do not exhaust the possibilities for data release, of course. Sometimes government agencies release summary information, such as a table, taken from more detailed data. These releases are neither microdata nor interactive data. *See* JACOB S. SIEGEL, APPLIED DEMOGRAPHY: APPLICATIONS TO BUSINESS, GOVERNMENT, LAW AND PUBLIC POLICY 175 (2002).

the disclosure risk expert prepares and releases microdata—individual-level datasets that researchers can download and analyze on their own. Other times, the expert prepares an interactive database that is searchable by the public. An outside researcher would submit a query or analysis request through a user interface that submits the query to the raw data. The interface returns the result to the outside researcher (sometimes after applying a privacy algorithm of some sort). The techniques for preserving privacy with these alternative research systems are quite different, not surprisingly. The debate over how best to prepare microdata is lively and rich.[18]

The public conversation about interactive databases, in contrast, is underdeveloped.[19] Outside of the technical field, hopeful faith in differential privacy dominates the discussion of query-based privacy.[20] This Part first explains the problem differential privacy seeks to solve. It is not immediately obvious why a query-based research system needs any protection for privacy in the first place, since outside researchers do not have direct access to the raw data; but even an interactive database can be exploited to expose a person's private information. Next, we demystify differential privacy—the creative solution developed by Microsoft researcher Cynthia Dwork—by working through a successful example of differential privacy in action.

A. The Problem

Six years ago, during a Eurostat work session on statistical data confidentiality in Manchester, England, Cynthia Dwork, an energetic and highly respected researcher at Microsoft, made a startling statement.[21] In a presentation to the world's statistical

---

18. One popular form of microdata release is the "de-identified" public database. De-identification involves the removal of all personally identifiable information and, sometimes, the removal of other categories of information that can identify a person in combination. HIPAA, for example, identifies 18 variables as personally identifiable information. 45 C.F.R. § 164.514(b)(2)(i)(A)–(R). Disclosure experts have long understood that de-identification cannot guarantee anonymization, but this subtlety is lost in news reporting. For a discussion of reidentification risk and its treatment in the popular press, see Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1, 36–37 (2011).

19. *Cf.* Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMMC'NS OF THE ACM 86, 89 (2011) (discussing the limited way the public uses interactive databases).

20. *See* Chin & Klinefelter, *supra* note 11, at 1452–53; Greengard, *supra* note 2, at 18; Ohm, *supra* note 10, at 1756–57; Wu, *supra* note 9, at 1137–38; Klarreich, *supra* note 7.

21. Cynthia Dwork, Presentation before the Eurostat Work Session on Statistical Data Confidentiality: Differentially Private Marginals Release with Mutual Consistency and Error Independent of Sample Size (Dec. 17–19, 2007), *available at* http://www.unece.org/fileadmin/DAM/stats/documents/2007/12/confidentiality/wp.19.e.ppt).

privacy researchers, Dwork announced that most, if not all, of the data privacy protection mechanisms currently in use were vulnerable to "blatant non-privacy."[22]

What Dwork meant by "blatant non-privacy" comes from a 2003 computer science publication by Irit Dinur and Kobbi Nissim.[23] Dinur and Nissim showed that an adversary—that is, a malicious false researcher who wishes to expose as much personal information as possible by querying a database—could reconstruct a binary database (a database containing only responses consisting of "0"s and "1"s) if they had limitless opportunity to query the original database, even if noise of magnitude $\pm E$ is added to the results of the queries, as long as E is not too large.[24] Dinur and Nissim defined "non-privacy" as a condition in which an adversary can accurately expose 99% of the original database through queries.[25]

To understand how such an attack works, suppose a database contains the HIV status of 400 patients at a particular clinic. The adversary knows that $E = 2$, meaning that the noise added or subtracted is no greater than 2. The adversary knows that for any response he receives from the system, the true value is within ± 2 of the response. Now assume that the adversary issues the query, "How many of the first 20 individuals in the database are HIV positive?" For the sake of argument, let us assume that the *true answer* to this query is 5. And assume that the system adds −2 to the true answer and responds with 3. Now the adversary asks: "How many of the first 21 individuals in the database are HIV positive?" Assume that the twenty-first individual is HIV positive, and the true answer to this query is 6. The system adds +2 to the true answer and responds with 8. From the response to the first query, the adversary knows that the true answer could not possibly be greater than 5. From the response to the second query, the adversary knows that the true answer could not possibly be less than 6. So, he can correctly conclude that: (a) the

22. *Id.* (emphasizing this point on slide 24 of the accompanying PowerPoint presentation); *see also* Cynthia Dwork, *Ask a Better Question, Get a Better Answer: A New Approach to Private Data Analysis*, *in* Database Theory – ICDT 2007: 11th International Conference 18, 18–20 (Thomas Schwentick & Dan Suciu eds., 2006) (describing the Dinur-Nissim "blatant non-privacy" vulnerabilities and proposing differential privacy as a solution).

23. Irit Dinur & Kobbi Nissim, *Revealing Information While Preserving Privacy*, *in* Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems 202, 204, 206 (2003).

24. To be precise, if the largest amount of noise added is E, and if E is less than the number of data subjects, Dinur and Nissim showed that an adversary who could make unlimited numbers of queries could reconstruct a database so that the new database differed from the old database in no more than $4E$ places. Thus, whenever $E < n/400$, the adversary will be able to construct a database that is accurate in 99% of the values, satisfying "blatant non-privacy." *Id.* at 205–07.

25. *Id.* at 204.

twenty-first individual must be HIV positive, and (b) there are 5 HIV positive cases among the first 20 individuals.

There are 2^{400} possible queries of this sort, and if an adversary used all of them, he could correctly reconstruct 99% of the HIV statuses. Dinur and Nissim also showed that even under more realistic scenarios where the number of queries is bounded, and even when the noise added occasionally exceeds E, an adversary can *still* recreate a rather accurate database as long as E is not too large and the value of E is known.[26]

These results provide important theoretical foundations for disclosure risk because they show that moving from a microdata release to a query system does not automatically assure privacy. A query system must be designed in a thoughtful way. However, from a practical perspective, the consequences of the Dinur-Nissim discovery are not as serious as they seem at first glance. For instance, if the selection of the noise function, E, is large enough, it can thwart an adversary's attempt to construct a nearly accurate database no matter how many queries he submits.[27]

But the most helpful limitation is the natural bound on the number of queries that a researcher can submit. Even for small databases, like the HIV database described above, an adversary would not be able to issue all of the queries necessary to attempt a full database reconstruction because of the sheer number of queries required. A database with 400 subjects would require 2^{400} queries. To give a sense of scale, $2^{332.2}$ is a googol, which is greater than the number of atoms in the observable universe.[28]

In addition to these natural limitations of the adversary, a query system may limit the total number of queries issued to the database or impose other restrictions when responding to queries.[29] The data producer can also withhold information about the amount of noise added. Once an adversary is constrained in the number of query submissions, an appropriate selection of noise can virtually guarantee

---

26. Conditioned on the fact that E is no larger than $\sqrt{n}$. *Id.* at 206.

27. For example, $E = 50$ would avoid blatant non-privacy for a small database with 1000 subjects because the reconstructed database would be off in $4 \times 50 = 200$ positions, rendering the database correct in only 80% of the values.

28. *See* John D. Cook, *There Isn't a Googol of Anything*, Endeavour (Oct. 13, 2010), http://www.johndcook.com/blog/2010/10/13/googol; *Googol*, Wolfram Math World, http://mathworld.wolfram.com/Googol.html (last visited Jan. 29, 2014) (discussing the size of a google).

29. For example, theoretically nothing prevents a researcher from querying "what is the HIV status of subject #2502?" *See* Klarreich, *supra* note 7 (noting that differentially private data release algorithms allow adversaries to ask "practically any question about a database," but "blur[s]" private information with noise).

that a reconstruction attack will not work.[30] The Dinur-Nissim attack would also fail if the administrator were to change the values in the original database and use the modified database to respond to all queries.[31]

Reconstruction attacks are not the only privacy threats that concern data providers. If an adversary can accurately figure out one highly sensitive attribute of a single data subject, such as an HIV diagnosis, the revelation would be disconcerting, even if the rest of the original database remained unknown. Meanwhile, data providers might shrug at a 99% accurate candidate database constructed by an "adversary" who guessed that everybody in the database had a negative HIV status.[32]

Thus, disclosure risk experts have long understood that the best approach to protecting privacy is one that is contextually sensitive.[33] Privacy risks fall disproportionately on data subjects whose demographics or other characteristics make them unusual.[34] Disclosure risk experts traditionally employ a range of techniques to protect outlier data subjects and highly sensitive attributes. Most of the time, for the sake of simplicity and ease of application, a database query system will add some random noise to the results generated by a particular query, and that noise usually falls within some bounded range.[35] That way, the utility of the response is not swamped by the noise added at the end. The disclosure limitation community was

30. Since a realistic adversary who is "bounded" or constrained by his computational ability will be thwarted by noise that is greater than $\sqrt{n}$, large databases require comparatively less noise to overcome the reconstruction attack. For example, a database with 100 subjects would require noise up to ±10 to avoid such an attack (10% of the total number of subjects), but a database with 1,000,000 requires noise only up to ±1000 (a tenth of one percent of the total number of subjects). *See* Dinur & Nissim, *supra* note 23, at 206.

31. Since the response to all queries are provided from the modified database, the best the adversary can hope to do is to reconstruct the modified database but not the original database.

32. A 99% accurate reconstruction is much more impressive when the binary outcomes are approximately equally likely (each outcome has probability approximately 50%). *See* Cynthia Dwork, *The Analytic Framework for Data: A Cryptographic View*, Microsoft Research 5 (2013), *available at* http://cusp.nyu.edu/wp-content/uploads/2013/06/chapter11v2.pdf.

33. *See* Tore Dalenius, *Towards a Methodology for Statistical Disclosure Control*, 5 STATISTISK TIDSKRIFT 429, 432–33 (1977) (explaining that the context of the data refers to "[t]he frame: $\{O\}_F$;" "[t]he data associated with the objects in the frame: *I*; *C*; *X*, *Y*, . . .,*Z*;" "[t]he statistics released from the survey: *S*;" and "[t]he extra-objective data: *E*" and noting that "[i]f the release of the statistics *S* makes it possible to determine the value D_K more accurately than is possible without access to *S*, a disclosure has taken place").

34. *See* Krishnamurty Muralidhar & Rathindra Sarathy, *Security of Random Data Perturbation Methods*, 24 ACM TRANSACTIONS ON DATABASE SYS. 487, 488 (1999); Rathindra Sarathy & Krishnamurty Muralidhar, *The Security of Confidential Numerical Data in Databases*, 13 Info. Sys. Res. 389, 393 (2002).

35. *See, e.g.*, Lawrence H. Cox & John A. George, *Controlled Rounding for Tables with Subtotals*, 20 ANNALS OPERATIONS RES. 141, 141 (1989); Dalenius, *supra* note 33, at 441.

interested in developing alternatives to these common noise-adding practices when Dwork made her provocative presentation.[36]

The holistic approach was unsatisfying to Dwork. She criticized the popular approaches for being "syntactic" and context driven.[37] Instead, Dwork insisted that the practical compromises were not necessary. One could design a query system that avoids even the *theoretical* risks of query attacks, or, rather, allows the theoretical risks only within a predefined range of tolerances.

B. The Birth of Differential Privacy

Differential privacy does two important things at once. First, it defines a measure of privacy, or rather, a measure of *disclosure*—the opposite of privacy.[38] And second, it allows data producers to set the bounds of how much disclosure they will allow.[39] For Dwork, if, based on a query result—or a series of results—an adversary can improve his prediction of a person's attributes, then *any such* improvement in the prediction represents a disclosure.[40]

In its purest form, this definition is too strong to be usable in settings where disclosure is strictly prohibited.[41] It obliterates research utility. Suppose, for example, an adversary has external knowledge that a particular person, Claire, is female. Now, any research describing gender differences along various dimensions would improve his predictions of Claire's attributes. While his best guess at her income would have been the average US income in the absence of better information, his prediction would be improved (though still not good) by learning that women earn less, on average, than men do. If disclosure were defined this broadly, every published statistic would violate privacy.

Dwork avoided this absurdity by proposing an elegant solution: differential privacy ensures that the presence or absence of an

---

36. *See, e.g.*, Muralidhar et al., *supra* note 15, at 1399; Muralidhar & Sarathy, *Data Shuffling*, *supra* note 15, at 658; D.B. Rubin, *Discussion of Statistical Disclosure Limitation*, 9 J. Official Stat. 461, 461 (1993).

37. Cynthia Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *in* PRIVACY, SECURITY, AND TRUST IN KDD–PINKDD 2007, at 1, 1 (F. Bonchi et al. eds., 2008).

38. *See id.* at 5–6; Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91; Cynthia Dwork, *Differential Privacy*, *in* 2 Proceedings of the 33rd International Conference on AUTOMATA, LANGUAGES AND PROGRAMMING 1, 8–9 (Michele Bugliesi et al. eds., 2006).

39. Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *supra* note 37, at 6; Dwork, *Differential Privacy*, *supra* note 38, at 9.

40. *See* Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *supra* note 37, at 6; Dwork, *Differential Privacy*, *supra* note 38, at 4.

41. *See* Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 89–90.

individual does not significantly affect the responses that the system provides.[42] More precisely, differential privacy disclosure occurs when, for any individual, the *probability* that a query will return a particular result in the presence of that individual in the database differs from the probability that a query would return that same result in the absence of that individual.[43] The *measure* of the disclosure for a particular query to a particular individual is the ratio of those two probabilities—the probabilities that the query system would return the result with, and then without, the individual's data.[44] Ideally, this ratio would be one, allowing no disclosure at all. But since this is impossible to achieve if the responses are to be useful, the data curator can select some small level of disclosure that society is willing to tolerate. The closer to one the ratio is, the less disclosure has taken place.[45]

For a query system to satisfy differential privacy, the system must add noise that ensures it only returns results such that the disclosure for everybody stays within certain predetermined bounds.[46]

Consider this example: Suppose a data producer had made differential privacy commitments, promising that the ratio of probabilities for all possible people and all possible values of return results would never be less than 1/2 or more than 2. And suppose that the database contains the wealth for the year 2010 for all Americans whose primary residence is in the state of Washington. An adversary submits the query, "How many people have more than $1 million in wealth?"

Suppose the true answer is 226,412, and one of those millionaires is Bill Gates.[47] The query system will apply some noise randomly drawn from a distribution, but what should that distribution be? Well, it must be drawn such that it does not diverge too greatly from the distribution of responses if the database *didn't* include Bill Gates. Removing Bill Gates from the database, the answer to the query is 226,411, and noise from the same distribution is randomly drawn to apply to that number instead. The query system must use a distribution that ensures that when we look at the probability of all possible returned results based on the true result or

42. *Id.*

43. *Id.* at 89.

44. *Id.*

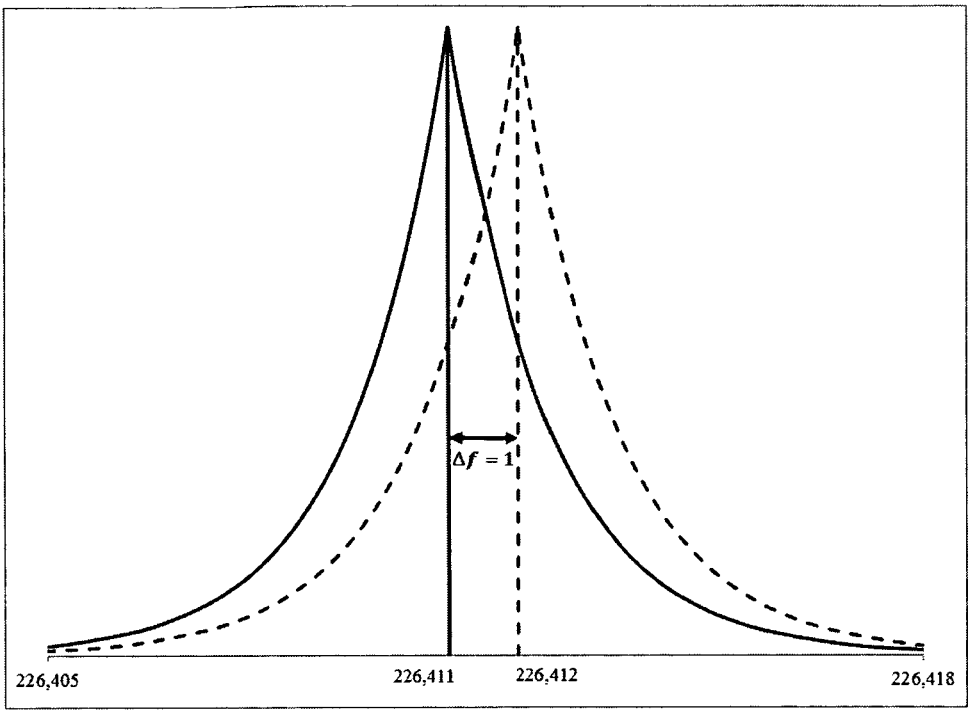
45. *See id.* at 87.

46. *See id.*

47. In 2010, the true figure was around 226,000. John Cook, *Millionaires to Double in Washington, but Will that Spark Angel Investment?*, GEEKWIRE (May 4, 2011, 2:04 PM), http://www.geekwire.com/2011/number-millionaires-double-washington-spark-angel-investment.

the result with a record deleted, the distributions are not too far apart. Figure 1 plots the distribution that has this quality.

Figure 1—Distribution of Query Response if the True Answer Contains, or Does Not Contain, Bill Gates



Reflect for a moment on the *reasons* that we want the query system to produce similar results whether Bill Gates is or is not in the query system. Most people know perfectly well that Bill Gates lives in Seattle and is a billionaire, so they would not be surprised to discover that he is included in the count of millionaires. But suppose an eccentric adversary knew the identity of every millionaire in Washington *except* Bill Gates. Suppose also that he knew that everybody except the 226,411 millionaires and Bill Gates were not millionaires. The only thing he does not know is whether Bill Gates has at least $1 million. If this adversary is clever, and if the data producer had used bounded noise, the adversary might be able to improve his inference that the noise centers around 226,411 (suggesting Gates is *not* a millionaire) or around 226,412 (suggesting that he is a millionaire).[48] Differential privacy ensures that the

48. For instance, the data producer may have added noise by selecting from random integer values in the range ±10. Hence, if the response to the query is 226,401, the adversary

system does not produce answers that behave very differently under either case.

Mathematically, the promise of differential privacy looks like this:

Given a database X, and a hypothetical database X^* that differs from X by the deletion or addition of just one record, differential privacy ensures that[49]

$$\frac{1}{e^{\varepsilon}} \leq \frac{P(Response = r|X)}{P(Response = r|X^*)} \leq e^{\varepsilon}$$

The data producer gets to choose ε, and the choice of ε will determine how much disclosure (as defined by Dwork and described above) the system will tolerate. The reason for the use of e (2.71828 . . .) is that by setting up the differential privacy promise this way, it corresponds precisely with a distribution curve already well known to statisticians—the Laplace distribution curve.[50] Laplace distribution has precisely the quality we are looking for: when the curve is shifted over a certain amount, the ratio of probabilities for the original and shifted curve stay within a predesignated boundary.

To employ differential privacy, a data curator would do the following:

(1) Select ε. The smaller the value, the greater the privacy.

(2) Compute the response to the query using the original data. Let a represent the true answer to the query.

(3) Compute the global sensitivity (Δf) for the query. Global sensitivity is determined by answering the following: "Assume that there are two databases X and X^* which differ in exactly one record and that the answer to this query from database X is a and that from database X^* is a^*. For any two such databases X and X^* *in the universe of all possible databases for the queried variable,* what is the maximum possible absolute difference between a and a^*?"[51] According

knows that Bill Gates is not a millionaire; if the response to the query is 226,422, the adversary knows that Bill Gates is a millionaire.

49. Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 90; Rathindra Sarathy & Krishnamurty Muralidhar, *Some Additional Insights on Applying Differential Privacy for Numeric Data*, *in* LECTURE NOTES IN COMPUTER SCIENCE: PRIVACY IN STATISTICAL DATABASES 210, 211 (Josep Domingo-Ferrer & Emmanouil Magkos eds., 2011) [hereinafter Sarathy & Muralidhar, *Additional Insights on Applying Differential Privacy*].

50. The probability density function of a Laplace random variable is $f(x) = \left(\frac{1}{b}\right) e^{-\frac{|x-\mu|}{b}}$.

51. In order to be able to compute Δf, a necessary step when implementing differential privacy, the data must have strict upper and lower bounds. Rathindra Sarathy & Krishnamurty Muralidhar, *Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data*, 4 TRANSACTIONS ON DATA PRIVACY 1, 4 (2011) [hereinafter Sarathy & Muralidhar, *Evaluating Laplace Noise*]; Sarathy & Muralidhar, *Additional Insights on Applying Differential Privacy*, *supra* note 49; Larry Wasserman & Shuheng Zhou, *A Statistical Framework for Differential*

to Dwork and Smith, "The sensitivity essentially captures how great a difference (between the value of $f$on two databases differing in a single element) must be hidden by the additive noise generated by the curator."[52] If the noise can protect this difference, then of course, all other, smaller, differences will also be protected. This is the key to differential privacy's protection.

(4) Generate a random value (noise) from a Laplace distribution with mean = 0 and scale parameter $b = \Delta f/\varepsilon$. Let y represent the randomly generated noise.

(5) Provide the user with response $R = a + y$. The noise added (y) is unrelated to the characteristics of the actual query (number of observations in the database or query and the value of the true response) and is determined exclusively by Δf and ε.[53]

Observe this as applied to the example of the number of millionaires in Washington. The data producer wanted the ratio of responses to stay within 1/2 and 2 when a person's information was included or removed from the database. Therefore, the data producer selected $\varepsilon = ln(2)$.[54] The global sensitivity here has to be one. Since the query asks for a headcount, the greatest difference any single person can make to the count is one.

We know that the true answer to the query is 226,412. We do not know what answer the data query system will produce because it takes the true answer and adds some randomly chosen noise from a Laplace distribution. But we can look at the range of responses such a system produces. Figure 2 plots the chance of seeing any particular response.

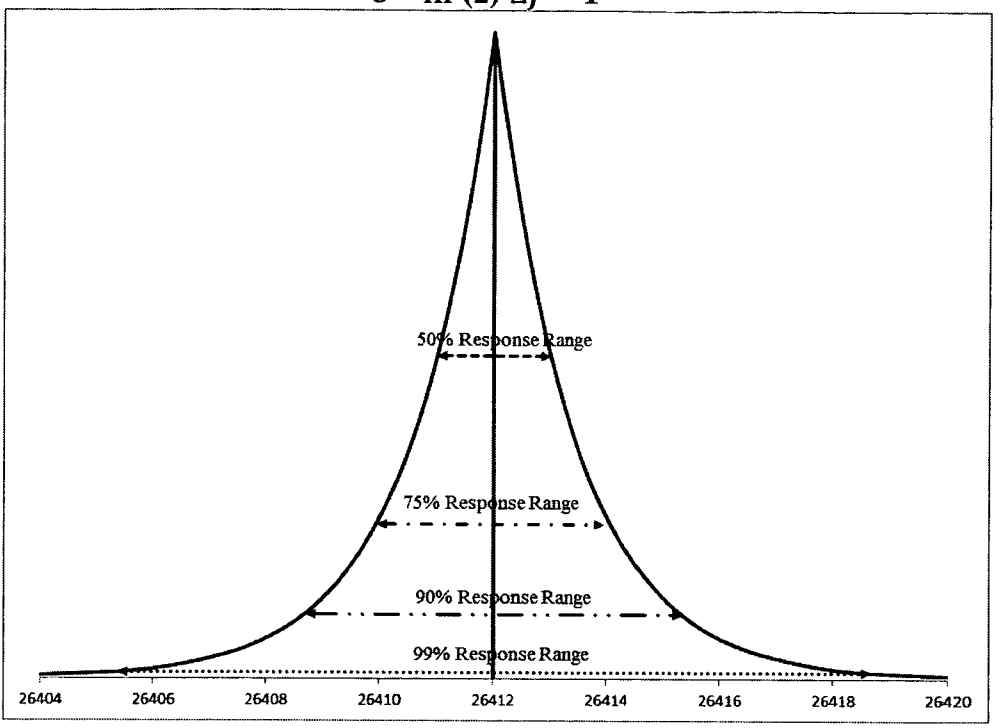
Privacy, 105 J. Am. Stat. Ass'n 375, 378–79 (2010) (noting that "it is difficult to extend differential privacy to unbounded domains").

52. Dwork & Smith, *supra* note 14, at 140.

53. "[O]ur expected error magnitude is constant, independent of n." Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92.

54. Surely you remember from precalculus class that $e^{\ln(2)} = 2$, right?

Figure 2—Number of Millionaires in Washington State
$\varepsilon = \ln(2)\ \Delta f = 1$



As you can see, differential privacy works quite well here. The query system produces results that tend to provide utility—the responses are very unlikely to be too far off from the true answer—and the system also insures against disclosure. This is a true win-win.

C. The Qualities of Differential Privacy

Much of this Article is devoted to illuminating the defects of differential privacy, but we do not want the reader to walk away without an understanding of its virtues. As the millionaires example demonstrates, Dwork's measure of disclosure makes the issue of auxiliary information easy to handle and potentially very privacy protecting. Even if the adversary knows everything in the database *except* one particular piece of information, differential privacy assures that the responses from the database—in the presence or absence of this record—are indistinguishable within a factor of e^{ε}. If we have confidence that this factor is small enough to be considered safe, then we need not speculate about what a user's motives are or how much information he already has. He can be a super-adversary, knowing almost everything, and his efforts will still be frustrated.

Differential privacy also protects against possible inferences based on a person's *absence* from a database.[55] A person's absence might reveal something very important. To see why this is so, return to the example of the income data for Washington residents. This time let us assume that the adversary's target is Larry Page, who does not live in Washington—and thus would not be in the database. If the last piece of information that the adversary needed about Larry Page was whether or not he lived in Washington, and the adversary also knew all of the 226,412 millionaires in Washington, then the fact that noise is not centered around 226,413 would reveal to the adversary that Larry Page does not live in Washington, and a disclosure would occur.

Dwork consciously made some overt choices and sacrifices when she developed differential privacy. For one thing, as Dwork herself has noted, microdata releases cannot be prepared in a way that strictly complies with differential privacy, so the standard applies only to query systems.[56] Also, much rides on the query designer's selection of ε. The smaller it is, the more privacy protecting, but also the more utility damaging since the noise added will tend to be larger.[57] Therefore, we must rely on the judgment of the data producer to select an appropriate ε that strikes the right bargain between privacy and utility.[58] This selection is all the more difficult because, whatever selection the data producer chooses for the system's overall privacy protections (ε), he must also decide how many queries researchers are allowed to make. Because the effects of successive queries on disclosure are cumulative, the data producer will have to divide his choice of ε by the anticipated number of queries.[59]

---

55. *See supra* notes 42–43 and accompanying text.

56. The definition of differential privacy "trivially rules out the subsample-and-release paradigm discussed: For an individual x not in the dataset, the probability that x's data is sampled and released is obviously zero; the multiplicative nature of the guarantee ensures that the same is true for an individual whose data *is* in the dataset." Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91. Thus, the very release of microdata violates DP requirements. In addition, the application of differential privacy is a function of the query submitted, and since microdata is released so that a person may use it to issue any and all queries, the promises of differential privacy cannot be kept. Sarathy & Muralidhar, *Evaluating Laplace Noise*, *supra* note 51, at 3. To meet the differential privacy standard, even if it were possible, the data producer would have to add so much noise that the database would be meaningless. Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92.

57. *See* Dwork, *A Firm Foundation for Private Data Analysis*, supra note 19, at 91–92.

58. As we will demonstrate later in this Article, a data curator who wants to preserve even a small amount of data utility will have to choose a fairly large ε, allowing a generous tolerance for disclosure. *See* discussion *infra* Part III.D.

59. To understand why this is so, let's revisit the Bill Gates example. The adversary knows that 226,411 individuals have more than a million dollars in personal wealth. Issuing the query "How many individuals in Washington State have more than a million dollars?" may result in a response that has twice the probability that the true answer is 226,411 compared to

Finally, in defining disclosure as she does, Dwork implicitly rejects other definitions of disclosure that would disclose families or groups.[60] Dwork ensures that an individual is not distinguishable from the results of a query, but she does not build in protections against revelations for families or subgroups.[61] What differential privacy *can* promise is that "the ability of an adversary to inflict harm (or good, for that matter)—of any sort, to any set of people—should be essentially the same, independent of whether any individual opts in to, or opts out of, the dataset."[62] For most research applications, this distinction between individuals and groups make sense.[63] After all, a research study finding that smoking causes cancer says something about every person who smokes—it allows an adversary to predict with better accuracy whether a particular smoker (whether they were in the research database or not) has cancer. But the adjustment to the adversary's prediction about that particular smoker would be based on group phenomena and not on individualized information about this particular smoker.[64]

Nevertheless, some data producers may be concerned about family and group disclosures. Some group disclosures—like whether a family has a congenital disease—might be more important than protecting against the theoretical possibility that somebody might not

---

the probability that the true answer is 226,412. The adversary can also issue the additional query "How many millionaires live in the 98039 zip code?," which happens to be Bill Gates's zip code. Jeanne Lang Jones, *The Sound's Wealthiest Zip Codes*, PUGET SOUND BUS. J. (Feb. 6, 2005, 9:00 PM), http://www.bizjournals.com/seattle/stories/2005/02/07/focus1.html. Since the adversary has information on all millionaires in Washington State, we have to assume that he also knows all the million dollar income earners (other than Bill Gates) who live in this zip code. The response to this query may result in a response that, as in the previous query, suggests that the probability that Bill Gates is a millionaire is twice as likely as Bill Gates not a millionaire. Since the Laplace noise has been added independently, taken together, these two results provide the adversary with the assurance that the probability Bill Gates is a millionaire is *four* times as likely as the probability that he is not a millionaire. The privacy specification for the two queries combined is thus $\ln(4) = 2 \times \ln(2) = 2\varepsilon$ (twice the original ε we had set). In general, if the adversary is allowed to issue m queries and the privacy assurance is set to ε for each query, then for all m queries combined, the privacy assurance is only $m\varepsilon$ (remember that a small ε provides more privacy). If we wish to limit the disclosure level to ε for all m queries combined, it would be necessary to set the disclosure level for each query to be (ε/m). Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92.

60. *See, e.g.*, Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 89.

61. *See infra* Part III.E.

62. Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91.

63. *See, e.g.*, Wu, *supra* note 9, at 1168–69.

64. Justin Brickell & Vitaly Shmatikov, *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*, *in* KDD '08 PROCEEDINGS OF THE 14TH ACM SIGKDD INT'L CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 70, 71 (2008) ("Sensitive attribute disclosure occurs when the adversary learns information about an individual's sensitive attribute(s). This form of privacy breach is different and in-comparable to learning whether an individual is included in the database, which is the focus of differential privacy."); *see also* Wu, *supra* note 9, at 1121–23 (further clarifying the difference between research-based and data-based disclosures).

know that Bill Gates lives in Washington. If so, they will have to rely on techniques beyond differential privacy.

II. STUNNING FAILURES IN APPLICATION

All database query systems serve the purpose of providing reasonably accurate information. Research results are the *raison d'être* for the query system in the first place. Inaccurate responses can be useless. In some cases, they can be positively harmful. Privacy is trivially easy to achieve if the data producer has no minimum standards for response accuracy. Responding to all queries with "0" would do the trick. Yet to facilitate useful research, maintaining reasonable accuracy has to be a priority. Unfortunately, differential privacy has great difficulty performing under most realistic conditions. The illustrations in this Part show that a data producer who wishes to comply with differential privacy will almost always have to choose between adding so much Laplace noise that the query results are ludicrous or adding so little noise that the dataset is left vulnerable to attack.

There are exceptions—the Washington millionaires example from the previous part is one of them. In Part III, this Article will explain when differential privacy can work. But first, let us examine how differential privacy can quickly go off the rails. As in most illustrations of differential privacy, we assume that the curator or administrator of the database allows for only one query to the database. This assumption is completely unrealistic since thousands (or perhaps millions) of queries may be issued to the database.[65] When the database receives many queries, the privacy afforded is diminished by each individual query.[66] We will consider this issue in more detail in Part III. The assumption of a single query presents differential privacy in the best possible light. Considering multiple queries means that the noise added will increase as a direct multiple of the number of queries, making matters much worse.[67]

---

65. *See* Drew Olanoff, *Zuckerberg on Building a Search Engine: Facebook Is Pretty Uniquely Positioned, at Some Point We'll Do It*, TECHCRUNCH (Sept. 11, 2012), http://techcrunch.com/2012/09/11/zuckerberg-we-have-a-team-working-on-search (stating that Facebook, for example, does over a billion queries a day).

66. *See* Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92 ("Given any query sequence $f_1, \ldots, f_m$, ε-differential privacy can be achieved by running K with noise distribution $\mathrm{Lap}\left(\sum_{i=1}^{m} \frac{\Delta f_i}{\varepsilon}\right)$ on *each* query, even if the queries are chosen adaptively, with each successive query depending on the answers to the previous queries.").

67. See *infra* Part III.D for a discussion of the multiple queries problem.

A. The Average Lithuanian Woman

One of the most frequently cited examples to justify the need for differential privacy is also, in our view, one of the most misguided. Dwork presents this example as she contemplates the disclosure risk from a database that includes the heights of Lithuanian women:

> Finally, suppose that one's true height is considered sensitive. Given the auxiliary information "[Alan] Turing is two inches taller than the average Lithuanian woman," access to the statistical database teaches Turing's height. In contrast, anyone without access to the database, knowing only the auxiliary information, learns much less about Turing's height. [68]

The idea is that even individuals who are not represented in the database stand to suffer a privacy violation.[69] Therefore, to set up the problem, we assume that (1) Alan Turing's height is not known to the public; (2) the height of the average Lithuanian woman is available only to those who have access to the query database; and (3) the auxiliary information that Turing is two inches taller than the average Lithuanian woman is known to the adversary.

This is an odd hypothetical. After all, in order to create the auxiliary information that "Turing is two inches taller than the average Lithuanian woman," the creator of the information must know both Turing's height *and* the height of the average Lithuanian woman. This would have to be Turing himself or somebody privy to his sensitive height information; but then, how did they know the height of Lithuanian women?

Even if a data curator is determined to protect height information, this particular style of auxiliary information falls outside the set of risks that differential privacy is designed to reduce.[70] The meat of the sensitive information is contained in the auxiliary information. The auxiliary information *is* the disclosure—it is just communicated in reference to some external fact.[71]

In any case, let us humor the hypothetical. What would differential privacy tell the curator of a database about the height of Lithuanian women to do in order to protect the privacy of Alan Turing—and others? Let us follow the steps laid out in Part I.

---

68. Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 90. The example has been repeated in other works, sometimes using Terry Gross instead of Alan Turing. *See, e.g.*, Dwork & Smith, *supra* note 14, at 136.

69. *See* Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 90–91.

70. *See* Wu, *supra* note 9, at 1137–38. Consider the following example. Suppose Turing declares: "My salary is ten times the zip code of the White House." Would publication of the White House's address violate Turing's privacy?

71. *See* Wu, *supra* note 9, at 1143–44. Felix Wu analogizes to the notions of cause-in-fact versus proximate cause. Disclosure of the external fact is a cause, but it is not a cause-in-fact. *Id.* at 1137–38.

1. Select ε

First, the curator of the database containing the height of Lithuanian women must decide on the value of ε (the acceptable level of disclosure). The curator must make a judgment call on how far off the probability distributions are allowed to be when the database does, and does not, include a particular person. Dwork has suggested that ε is often in the order of 0.01 or 0.1, "or in some cases, ln 2 or ln 3."[72] Since the primary objective in this exercise is to prevent disclosure, we should use a fairly high privacy standard, setting $\varepsilon = 0.1$. (Remember, the smaller the ε, the greater the noise).

The query "What is the height of the average Lithuanian woman" is actually two queries rolled into one because it requires two different pieces of information: the number of Lithuanian women and their total height. Further, since $\varepsilon = 0.1$ and the response involves two different queries, for each query, we will set $\varepsilon_q = 0.05$.

2. Compute the Response to the Query Using the Original Data

According to *Statistics Lithuania*, the population of Lithuania in 2012 was just over 3 million, with females accounting for approximately 1.6 million.[73] The average height of Lithuanian women is 66 inches.[74]

3. Compute the Global Sensitivity (Δf) for the Query

We must determine global sensitivity for both the count of Lithuanian women and the sum of their heights. The absence or presence of an individual will change the number of Lithuanian woman by exactly one and hence $\Delta f = 1$. But how about the sum of the height query? The largest difference in the sum of heights between any two databases that differ in one record would occur when one database contains the tallest living person and the other does not. The difference in the total height between the two databases would equal the height of the tallest living person. The height of the tallest person living in the world today is 99 inches (8'3"), so Δf for the sum of the height query is 99.

72. Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91.

73. *See Official Statistics Portal*, Stat. Lith. (Apr. 9, 2014), http://osp.stat.gov.lt/en/temines-lenteles19.

74. *See Average Female Height by Country*, AVERAGEHEIGHT.CO, http://www.averageheight.co/average-female-height-by-country (last visited Feb. 5, 2014).

4. Generate a Random Value (Noise) from a Laplace Distribution with Mean = 0 and Scale Parameter $b = \Delta f/\varepsilon$

Based on the information worked out above, the Table provides the original answers, the noise added, and the response to a query operating on the entire population of Lithuanian women.

Table 1—Response to Query on Average Height Over Database of Lithuanian Women

$\varepsilon_q = 0.05$

	True values	Δf	Laplace Noise		Noise Added Response	
			Low (0.01)	High (0.99)	Low	High
# of Lithuanian Women	1,603,014	1	−78	78	1,602,936	1,603,092
Total Height (inches)	105,798,924	99	−7,746	7,746	105,791,178	105,806,670
Average Height (inches)	66				65.99	66.01

Because this query analyzes over one million people, the large n keeps the Laplace noise from drowning out the true signal. Thus, the low estimate of average height is within 0.02" of the high estimate for average height. Anyone who knows that Turing is 2" taller than the average Lithuanian woman will have no trouble concluding that he is 68" tall, even after the data curator adopts the precautions of differential privacy.

However, the decision to adopt differential privacy to protect *everyone* (including Turing and the world's tallest person), *whether or not they are in the database*, comes at a very high cost in other contexts. What if the adversary knew that Turing was 2" taller than the average woman in the small Lithuanian town of Smalininkai (population 621, of whom 350 are women)? Or what if the adversary knows Turing is 2" taller than the average *employed* woman in Smalininkai? Now, to protect the possibility of disclosure for Turing (as well as the world's tallest person), the query system must allow the possibility of inventing a land of 30-foot-tall women. It also may produce tiny towns with people measuring less than 1" tall. Tables 2 and 3 display the range of results for average heights of these smaller subpopulations, using the same differential privacy parameters we set before.

Table 2—Response to Query on Average Height of Smalininkai Women Over Database of Lithuanian Women $\varepsilon_q = 0.05$

	True values	Δf	Laplace Noise		Noise Added Response	
			Low (0.01)	High (0.99)	Low	High
# of Smalininkai Women	350	1	−78	78	272	428
Total Height (inches)	23,100	99	−7,746	7,746	15,354	30,846
Average Height (inches)	66				35.9	113.5

Table 3—Response to Query on Average Height of Employed Smalininkai Women Over Database of Lithuanian Women $\varepsilon_q = 0.05$

	True values	Δf	Laplace Noise		Noise Added Response	
			Low (0.01)	High (0.99)	Low	High
# of Employed Smalininkai Women	120	1	−78	78	42	198
Total Height (inches)	7,920	99	−7,746	7,746	174	15,666
Average Height (inches)	66				0.88	375.1

Notice that the distributions of noise that the equation adds to the count and total heights in Tables 2 and 3 are identical to the distributions shown in Table 1. This should not be surprising, since the shape of the noise distribution is determined solely by the values of ε_q and Δf. These values did not change since we still have to protect the world's tallest person. However, while the noise was relatively small as applied to the entire female population of Lithuania, the same noise quickly overwhelms the true values when taking the averages over smaller subpopulations.

One could rationalize that smaller subgroups need more noise to protect the confidential information. However, research databases often rely on randomly selected subsamples of the population to avoid the significant costs of surveying every person. The database applies the exact same distribution of noise to an unknown, random

subsample of the population. So, if a world census allowed researchers to query average heights on a randomly selected sample of 120 Lithuanian women, the results would look just as bizarre as the ones reported in Table 3.

Matters would be much worse if we assume that the curator decides to respond to several hundred or thousands of queries. The noise currently added is large enough to overwhelm the true answer; with one thousand queries, the noise added to comply with differential privacy standards would increase a thousand fold![75]

B. Averages of Variables With Long Tails

Differential privacy has the potential to radically distort averages of variables (like height) that are normally distributed, but the distortion is even worse on variables like income that have a skew—that is, where some members of the population have values that are very distant from the median. For instance, while the median family income in the United States is just under $53,000,[76] a few hedge fund operators like George Soros have income exceeding $1 billion.[77] Scholars often refer to these distant values to as the "long tail" of the distribution.

Booneville, Kentucky, is a small and struggling town.[78] Its population is just over 100, and the median household income is just above the poverty line.[79] Suppose the town decided to make a database available for public research as part of a new transparency initiative designed to inspire research on public welfare and the prevention of poverty. Under normal circumstances, one might counsel the town to include only a random subsample of residents and to join forces with other similar towns so that a data user might not be able to discern the precise town in which the data subjects live. There may be other precautions too, based on the context and nature of the data. But in this hypothetical scenario, the town has opted instead to rely on differential privacy. After all, one of the core strengths of

---

75. See *infra* Part III.D for a discussion of the queries problem.

76. *Selected Economic Characteristics: 2007–2011 American Community Survey 5-Year Estimates*, US Census Bureau, http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_5YR_DP03 (last visited Jan. 26, 2014).

77. Louise Story, *Top Hedge Fund Managers Do Well in a Down Year*, N.Y. TIMES, Mar. 24, 2009, http://www.nytimes.com/2009/03/25/business/25hedge.html.

78. *See Selected Economic Characteristics: Booneville City, Kentucky, 2007–2011 American Community Survey 5-Year Estimates*, US Census Bureau, http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml (search for "American Community Survey" and "Booneville city, Arkansas"; then show results from 2011) (last visited Feb. 5, 2014).

79. *See id.*

differential privacy is that the methods of masking query responses are completely independent of the size and nature of the Booneville data—the town can have mathematical certainty of meeting privacy standards regardless of the particular features of its town.[80]

What happens when a researcher queries the average income of Booneville residents? In this case, income is the confidential variable; we do not want an adversary to be able to tell something about his target—either about his income or using his income—based on what he learns from the response to the query. In particular, the town would need to ensure that the adversary would not be able to rule out that his target—a Booneville resident—is a billionaire. After all, when large values are included in an analysis of the mean, the outlier has an outsized effect on the analysis. So a reported mean that roughly matches the incomes of the rest of the Booneville population would suggest that the last person in the sample is not a billionaire. Also, the town might need to ensure that an adversary who knows everything about George Soros except where he lives is not able to rule out Booneville as George Soros's hometown. Thus, even if the highest income among Booneville residents is $50,000, the probability of any particular response coming back from the query needs to be not so far off from the probability that that response would come back if George Soros lived in Booneville.[81] That is the promise of differential privacy. Unfortunately, this privacy promise also means that the response is likely to be useless.

Now, we will work through the application following the instructions we provided in Part I.

1. Select ε

First, the town must decide how much disclosure it is willing to tolerate and will have to allocate this disclosure among all the queries it issues to this database. For simplicity we will assume that the town will use $\varepsilon = 0.50$ for this particular query.[82]

80. *See* Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91.

81. The fact that everyone knows with practical certainty that no one in the subset earned $1 billion is irrelevant; the response distribution should be constructed in such a manner that $1 billion income is feasible in this subset. *See* Dwork & Smith, *supra* note 14, at 137.

82. Note that this selection is less differential privacy-protecting, and thus more utility-preserving, than our last example.

2. Compute the Response to the Query Using the Original Data

Suppose, for this illustration, the true per capita income for Booneville residents is $23,426 (which is the value reported by the US Census Bureau's FactFinder web tool for 2007–11).[83]

3. Compute the Global Sensitivity (Δf) for the Query

As we saw with the example of Lithuanian women, this query actually involves two separate global sensitivities (sum of income and count of people), but we will take a shortcut by dividing the global sensitivity for income by the number of data subjects responsive to the query.[84] In this case, only 59 Booneville residents were in the workforce according to FactFinder.[85]

When it comes to income, the global sensitivity is very large. It is the difference between the highest-paid man in the world and an unemployed man. For the sake of illustration, we will assume that the highest income is $1 billion and the lowest is $0. Thus, the global sensitivity is $1 billion.[86]

4. Generate a Random Value (Noise) from a Laplace Distribution with Mean = 0 and Scale Parameter $b = \Delta f/\varepsilon$

Now comes the fun part—the selection of noise to add to the true answer ($23,426). A Laplace distribution randomly selects noise, but the reason we went through all the work of determining the global sensitivity and the value of ε is that these two factors determine the distribution—the likelihood of how much noise the equation adds. To satisfy differential privacy, the Laplace distribution which randomly selects the noise must have a standard deviation of $\sqrt{2\frac{\Delta f}{n\varepsilon_q}} = \sqrt{2\frac{1000000000}{59\times 0.5}} \approx 48$ million.

Thus, although the true answer to the query "What is the average income of the inhabitants of Booneville?" is $23,426, the answer after the differential privacy process is very likely to be over

83. *See Selected Economic Characteristics: Booneville City, Kentucky*, *supra* note 78.

84. For the purposes of this illustration, we have added noise only to the income variable. Adding noise to the number of residents would have made matters worse.

85. *See Selected Economic Characteristics: 2007–2011 American Community Survey 5-Year Estimates*, *supra* note 76.

86. We know that hedge fund operators like George Soros regularly take pay in excess of $1 billion, so our illustration is a *conservative* estimate of the noise that would be added by differential privacy processes.

$10 million.[87] It is also very likely to come out lower than negative $10 million. In fact, the chance that the query answer will be within $1 million of the true answer is under 3%.[88]

Table 4 and Figure 3 show the Laplace distribution of noise. The two dotted lines represent negative $5 million and $5 million. The small area between the dotted lines visually represents the chance that the noise would fall within that range.

Table 4—Distribution of Noise Added to a Query for Average Income Where the True Answer is $23,426
$\varepsilon_q = 0.5$, Δf = $1 Billion

Noise Level	Noise Added	Response (True Value + Noise)
Very Low (0.001)	−210,664,681	−$210,641,255
First percentile (0.01)	−132,610,949	−$132,587,523
Fifth percentile (0.05)	−78,053,732	−$78,030,306
Tenth percentile (0.10)	−54,557,217	−$54,533,791
Twenty-fifth (0.25)	−23,496,515	−$23,473,089
Fiftieth (0.50)	0	$23,426
Seventy-fifth (0.75)	23,496,515	$23,519,941
Ninetieth (0.90)	54,557,217	$54,580,643
Ninety-fifth (0.95)	78,053,732	$78,077,158
Ninety-ninth (0.99)	132,610,949	$132,634,375
Very High (0.999)	210,664,681	$210,688,107

87. *See* Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *supra* note 37, at 7.
88. *See id.* at 8.

Figure 3—Distribution of Noise Added to a Query for Average Income $\varepsilon_q = 0.5$, Δf = \$1 Billion

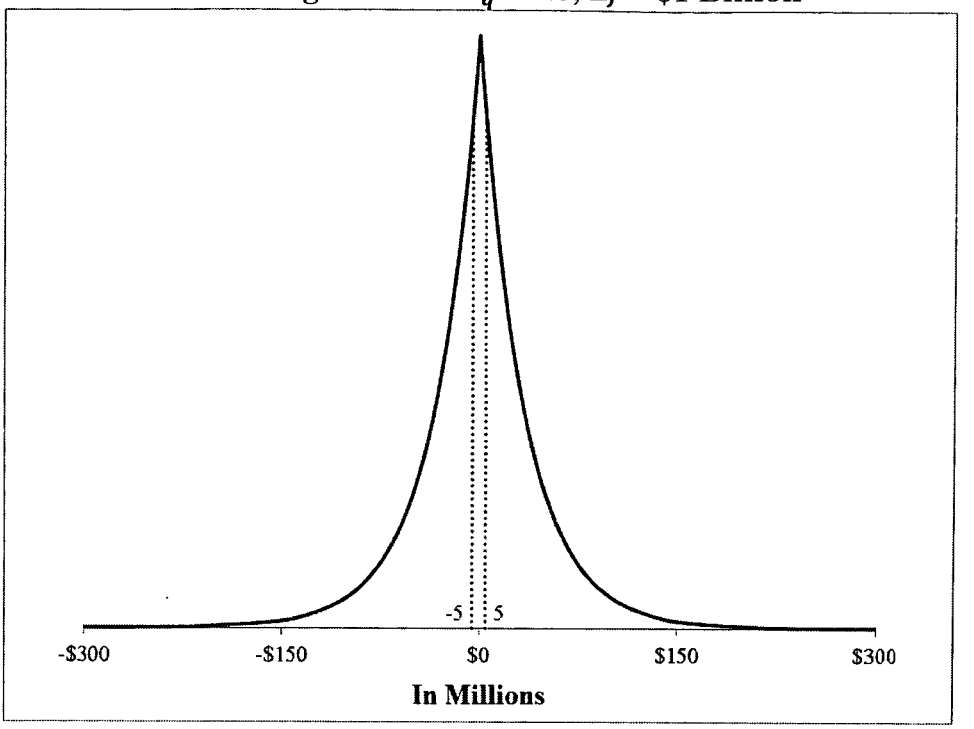


Table 5 shows the distribution of noise under various choices ofε. Even if the data producer chose 1 for the value of ε, a choice that might garner criticism for being insufficiently protective of privacy, the response to any query on the income variable would be swamped by noise.

Table 5—The Probability that Laplace Noise Will Be Selected from Specified Ranges Δf = \$1 Billion

	$\varepsilon = 0.01$	$\varepsilon = 0.10$	$\varepsilon = 0.50$	$\varepsilon = 1.00$	$\varepsilon = \ln(3)$
±10,000	0.0000	0.0001	0.0003	0.0006	0.0006
±100,000	0.0001	0.0006	0.0029	0.0059	0.0065
±500,000	0.0003	0.0029	0.0146	0.0291	0.0319
±1 Million	0.0006	0.0059	0.0291	0.0573	0.0628
±5 Million	0.0029	0.0291	0.1371	0.2555	0.2768
±10 Million	0.0059	0.0573	0.2555	0.4457	0.4770
±100 Million	0.2555	0.9477	1.0000	1.0000	1.0000
±1 Billion	0.4457	0.9973	1.0000	1.0000	1.0000

Table 5 also reveals another important fact about differential privacy method; by design, the noise added to a query is entirely independent from the values of the database. The Laplace noise

distribution is determined by global sensitivity and the choice of ε, neither of which required the data producer to consult the database.[89] The noise is independent from the actual answer to the query.[90] So Table 5 represents the noise that would be added not only to this hypothetical query involving a small town in Kentucky but to *any* analysis of income over data this size. Therefore, if the US Census Bureau chose to adopt differential privacy in an online query system for the Current Population Survey, it too would add and subtract hundreds of millions in noise to protect George Soros when a user queried, "What is the average income for employed females over the age of 65 living in the South Bronx?" Note that this applies even to queries about females because the last pieces of information an adversary might need about George Soros is that he is *not* an older female living in the Bronx.

When it comes to the analysis of continuous, skewed variables (like income), differential privacy's strict and inflexible promises force a data producer to select from two choices: he can either obliterate the data's utility or he can give up on the type of privacy that differential privacy promises.

For comparison's sake, let us look at how the Census Bureau's American FactFinder service actually reports the income of the residents in Booneville, Kentucky.[91] According to American FactFinder, the average income of the 51 working individuals in Booneville is $21,907 and a margin of error of ±$11,247.[92] For any realistic selection of ε, this release of information by the Census Bureau would violate differential privacy since an adversary would be able to conclude that it is extremely unlikely that anyone living in Booneville has an income of $1 billion. From the first line of Table 5 above, one can see that the probability of observing a differentially private response within the range that the Census Bureau has released is infinitesimally small.

It is hard to fault the Census Bureau for not using differential privacy. After all, a little external information and knowledge of the world would suggest that it is extremely unlikely that a multi-billionaire lives in a small, poor town in Kentucky. It makes little sense to guard against the revelation that, as one would expect, there are no billionaires in Booneville at the cost of the utility of the rest of the dataset. Differential privacy does not differentiate between the

89. "Thus, our expected error magnitude is constant, independent of n." Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92.

90. *See id.*

91. *See Selected Economic Characteristics: Booneville City, Kentucky*, *supra* note 78.

92. *Id.*

many possible types of revelations. It treats all as if they were equally meaningful, which leads to silly results and upside-down priorities.

C. Tables

Part I demonstrated that differential privacy can perform fairly well when queries are asked to report counts, such as the numbers of people who have various characteristics. Suppose that, instead of querying the mean income, the data user submitted a query to create a histogram of income? With count queries, the addition or deletion of one individual changes only a single bucket in a histogram—and by only 1. Thus, the global sensitivity is 1 instead of $1 billion.

Before we present the results, it is worth reflecting on the loss of utility that comes with the change of format. The accuracy of simple statistics from grouped histogram data is always compromised by the crudeness of the categories. Still, one might expect an improvement over the differential privacy responses for average income that we explored above.

Table 6 shows a hypothetical histogram for Booneville, Kentucky, and noise that we randomly selected from a Laplace distribution with $\varepsilon = 0.50$ (as before). This is just one realization of possible responses to the histogram query. In practice, the data user would see only the last column of the table. The shaded columns help us assess whether the last column is close enough for research purposes.

Table 6—Example Responses to a Series of Count Queries about the Income of Booneville Residents $\varepsilon_q = 0.5$, $\Delta f = 1$

Income Group	True Count	Noise (rounded to the closest integer)	Response (True Count + Noise)
$0 to $10 Thousand	11	2	13
$10 Thousand to $50 Thousand	40	7	47
$50 Thousand to $100 Thousand	7	−2	5
$100 Thousand to $500 Thousand	1	−4	−3
$500 Thousand to $1 Million	0	−5	−5
$1 Million to $10 Million	0	0	0
$10 Million to $100 Million	0	3	3
$100 Million to $1 Billion	0	0	0
More than $1 Billion	0	5	5

The unshaded response column reports that there are five individuals whose income is higher than $1 billion and three individuals whose income is between $10 million and $100 million. Of course, we know that the maximum income of individuals in Booneville city is less than $500,000, so this table steers researchers wildly off the mark.[93] Naturally, the negative values are par for the course.[94] They very slightly help balance out the bias from positive noise if the researcher decides to use the table to calculate a rough estimate of average income, but the correction is hardly worth the bother since an estimate of the average would be quite poor as it is. A researcher using only the responses above would conclude that the average income among Booneville residents is about $44 million.[95]

Why does this table perform so poorly even though the table from Part I, reporting the number of millionaires in Washington, performed so well? Recall that the noise or, more precisely, the distribution that produces the noise, is independent from the true values in the original dataset. It is also independent from the size of the database. In both tables, the global sensitivity (Δf) is 1. However, when working with the number of Washington millionaires, noise in the range of −7 to 7 does not make much of a difference because the true response is over 200,000. Here, since the true answers are small (under 100), noise on the same scale greatly distorts the analysis.

Table 7 shows the Laplace distributions for tabular data, where $\Delta f = 1$. Each row displays the probability of observing noise values within the identified range for varying specifications of ε.

93. One option for skewed data is to set arbitrary upper and lower limits for the values. For the income variable, it might be suggested that the upper limit should be set at (say) $100 thousand. For this particular query, such a truncation would eliminate the problem of very large values. But the truncation would frustrate research on high income earners, or on income inequality. For example, if the query asked for the average income of hedge fund managers, truncating the upper limit of income at $100 thousand would put nearly the entire data set in the truncated range. *See* J.K. Ord et al., *Truncated Distributions and Measures of Income Inequality*, 45 INDIAN J. STAT. 413, 414–15 (1983).

94. *See* Microsoft, *supra* note 4, at 5.

95. Assuming that the researcher sets the income in the middle of the range for each category, so that the 23 people earning between $0 and $10,000 are estimated to earn $5,000, the 85 people earning between $10,000 and $50,000 are estimated to earn $30,000, etc. The 5 people earning in excess of $1 billion are estimated to earn $1 billion and $1. By this method, the researcher would reach an estimated average income over $44 million. Using the same message using the "True Count" column would yield a more modest average income of $35,254. We know that this is still quite far from the $21,907 average that the Census reports for the town. *See Selected Economic Characteristics: Booneville City, Kentucky*, *supra* note 78.

Table 7—The Probability that Laplace Noise Will Be Selected from Specified Ranges, for Varying Selections of ε $\Delta f = 1$

	0.001	**0.01**	**0.10**	**0.25**	**0.50**	***ln*(2)**	**1.00**	***ln*(3)**	**5.00**
±1	0.00	0.01	0.10	0.22	0.39	0.50	0.63	0.67	0.99
±2	0.00	0.02	0.18	0.39	0.63	0.75	0.86	0.89	1.00
±3	0.00	0.03	0.26	0.53	0.78	0.88	0.95	0.96	
±5	0.00	0.05	0.39	0.71	0.92	0.97	0.99	1.00	
±10	0.01	0.10	0.63	0.92	0.99	1.00	1.00		
±20	0.02	0.18	0.86	0.99	1.00				
±50	0.05	0.39	0.99	1.00					
±100	0.10	0.63	1.00						
±500	0.39	0.99							
±1000	0.63	1.00							
±5000	0.99								
±10000	1.00								

When $\varepsilon > 1$, relatively little noise is added to the true answer. But, large ε values open the system to risk of disclosure, and the risk is not managed in any thoughtful way. When ε is as large as 5 or higher, the risk of disclosure is so great that the system cannot fairly be described as a privacy-protecting one. When $\varepsilon < 0.10$, the noise generated could be ±100. Adding 100 or more to a query response might be just fine if the true response is in the order of 100,000 or more, but it causes chaos if the true answer is less than ten. Table 7 shows the distribution of noise added to count queries *irrespective* of the true answer. Once ε is specified, the noise will be generated with the above stated probabilities.

Dwork defends this as a desirable feature since small databases leave the data subjects more vulnerable and thus require proportionally more protection than larger databases.[96] But this is not necessarily so. Suppose that Table 6, the representative example of a histogram query, reports the income not from the town of Booneville, but from a stratified random sample of 130 Americans. As long as the adversary does not have a way of knowing who was included in the random sample, this database would not require any more protective noise than a database containing the entire US population, yet

96. *See* Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91.

differential privacy methods would cause much more loss to its utility.[97]

Moreover, the noise distribution is not limited to count queries. This noise is added in *all situations* for which $\Delta f = 1$, even if the query demands a strict upper and lower bound for the true value. Consider the query, "What is the average income tax rate for Americans?" A person submitting the query would expect a reasonable response between 0% and 39.6% (the highest marginal tax rate), but Table 7 shows that for any $\varepsilon < 5.0$, there is a high probability that the response will be negative or above 1, rendering it useless. This is also poses a significant problem for statistical measures that must be interpreted within a bounded range, as we illustrate in the next example.

D. Correlations

Lest there be any doubt that differential privacy performs poorly under most typical research settings, consider its effects on correlation. Statistical research often explores the relationships between variables. Pearson's product-moment correlation, measuring the strength of the linear relationship between two variables, is one of the most basic and essential tools to understand how various forces and phenomena interact and operate on one another. Correlation ranges between [−1, 1] where −1 means that two variables have a perfectly negative relationship (an increase in X corresponds with a proportional decrease in Y), 0 means the two variables share no relationship (an increase in X sometimes corresponds with increases and sometimes decreases in Y), and 1 indicates a perfectly positive relationship (an increase in X corresponds with a proportional increase in Y). In this case, the function (correlation) has clear lower and upper bounds—a query on correlation will *always* come out between −1 and 1.

Suppose the Department of Education is preparing a database query system based on a national longitudinal study on the relationship between education and income. Among other things, the database contains information on each data subject's highest educational attainment (measured in years of qualified schooling) and annual income. What happens when the Department of Education adopts differential privacy and applies Laplace noise to a query

97. It also seems to contradict the work of Dinur and Nissim, who conclude that in order to prevent blatant non-privacy, the noise added would have to be in the order of $\sqrt{n}$. Dinur & Nissim, *supra* note 23, at 206.

requesting the correlation between educational attainment and income?

Let us work through the usual steps:

1. Select ε

In this example, let us explore what happens to the query response under a range of ε running from 0.01 (relatively privacy protective) to 10.0 (quite lax). As before, we will assume a single query of the database to avoid the need to add more noise for serial queries.

2. Compute the Response to the Query Using the Original Data

The relationship between education and income is strong. Expected earnings increase in lockstep as a person moves from high school to college to masters, doctoral, or professional degrees.[98] Assume for this exercise that the education and income data in the Department of Education's database produce a correlation coefficient of 0.45.

3. Compute the Global Sensitivity (Δf) for the Query

The global sensitivity requires the data curator to anticipate the greatest difference that the addition or subtraction of a single data point can make to a similar query on the same variable for *any* possible database—not just the database that the curator is preparing for public research.[99]

For a very small sample, the addition (or subtraction) of a single data subject can change the correlation coefficient of two variables from perfectly positive correlation to a strong negative correlation, or vice versa—a change of nearly 2. To see how, imagine a database with just two people. Person A has had fewer than 8 years of formal education (no high school) and has an annual income of $52,000. Person B has a professional degree and earns $70,000 each year. For this small set of data, correlation between education and income will be 1: the more education, the more income. Now, imagine what happens when we add Person C to the dataset. Person C also has no formal education, but has an income of $1 million. With these three data points, the correlation between income and education can

98. Sandy Baum & Jennifer Ma, *Education Pays: The Benefits of Higher Education for Individuals and Society*, COLLEGE BOARD RESEARCH PAPER 10 (2007).

99. *See* Part I.B (discussing the need for the data curator to mask the presence or absence of any entry).

fall below 0. After adding Person C, it looks like on balance, less education will tend to increase income.

We could construct a similar illustration where a correlation of +1 is converted to −1 (or something infinitely close) with the addition of 1 new data point, so we are working with $\Delta f = 2$.

4. Generate a Random Value (Noise) from a Laplace Distribution with Mean = 0 and Scale Parameter $b = \Delta f/\varepsilon$

Next we randomly draw noise from the Laplace distribution determined by the values of global sensitivity and ε. This is where the process takes a turn for the worst.

Correlation takes the range from −1 to 1. Output outside of that range would be meaningless, and small changes within the range can have a great effect on the researcher's interpretation. Table 8 reports the probability that the noise added to the true answer will be no higher than 1, and no lower than −1 under varying selections of ε.

Table 8—The Probability that Laplace Noise Will Fall Within [−1, 1] for Varying Selections of ε

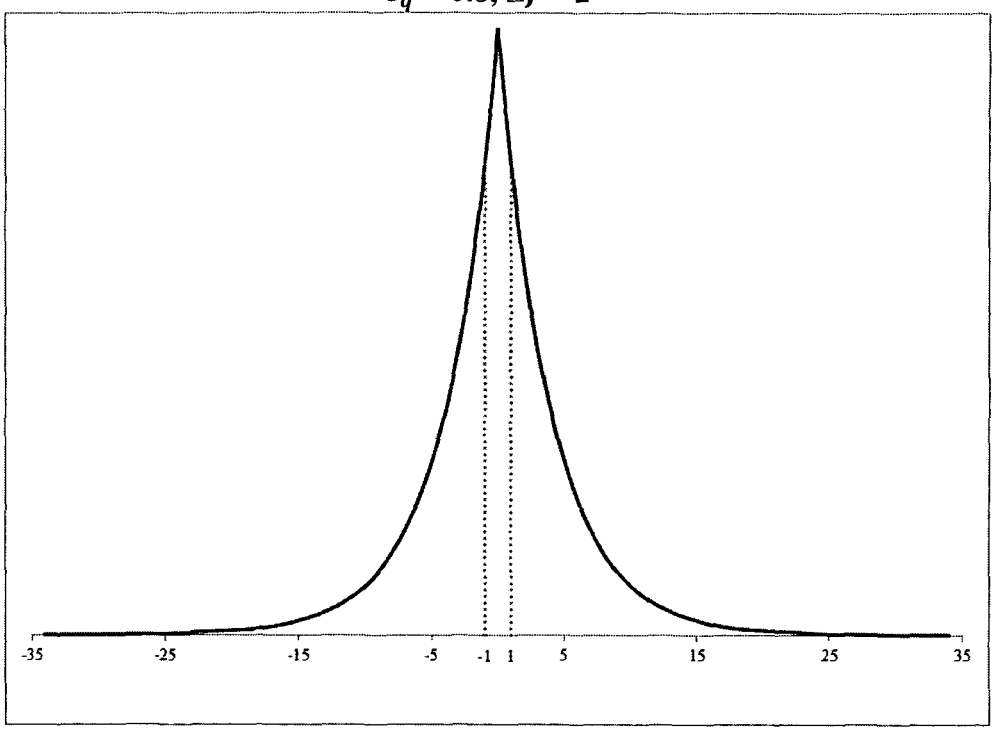
$\Delta f = 2$

ε	Probability Noise Is in the Range [−1, 1]
0.01	0.004988
0.10	0.048771
0.20	0.095163
0.50	0.221199
1.00	0.393469
2.00	0.632121
5.00	0.917915
10.00	0.993262

For small, privacy-protecting levels of ε (< 0.50), the noise added to the true answer is very likely to be so large that the query system's response will be nonsense. If the data curator selects $\varepsilon \geq 5$, there is a decent chance the reported correlation will be within the range, but of course it is also very likely to misstate the relationship between the variables (and to say that two factors that are positively correlated are negatively correlated, or vice versa).

Figure 4 shows what the distribution of responses would be, assuming that the true answer (the actual correlation) is zero and ε = 0.50. The dotted lines show the acceptable response range [−1, 1]. The figure illustrates that the great majority of responses would fall outside the acceptable range for correlation rendering the response completely meaningless to the user. Many of the responses *within* the dotted lines would be very misleading to the researchers and to the relying public.

Figure 4—Distribution of Responses to a Query for Correlation Where the True Answer is 0
$\varepsilon_q = 0.5$, $\Delta f = 2$



With noise like this, differential privacy simply cannot provide a workable solution for analyses of correlations or of any statistical measure with a strict upper and lower bound.

The examples worked through in this Part should give a sense of differential privacy's serious practical limitations. While differential privacy is a technical standard, the problems that it would cause if adopted broadly would be profound, wide reaching, and devastating to research. Nevertheless, policymakers and privacy

scholars are embracing differential privacy with increasing enthusiasm.[100] This enthusiasm must be tempered. The proponents of differential privacy have oversold its usefulness. Realistically, the future of data privacy will rely on differential privacy only in very narrow circumstances or only if differential privacy is modified to the point of being unrecognizable to its creators.

III. THE GOLDEN HAMMER

The proponents of differential privacy have embraced the law of the instrument: When you have a hammer, everything looks like a nail. The developers of differential privacy have insisted that it is a full-service tool that will free research from the perils of privacy risk in every context. As Cynthia Dwork and her collaborators say, apply differential privacy "and never look back."[101]

Policymakers and legal scholars are ready to adopt differential privacy as a—or even *the*—best practice, though their enthusiasm reveals a lack of understanding about what differential privacy would do to data research.[102] In one case, legal scholars jumped to the conclusion that Facebook employs differential privacy when it is very likely using a different noise-adding technique.[103] This is a variation on the law of the instrument: When you like hammers, every tool looks like one.

In this Part, we will explore why differential privacy has suddenly gained the attention and trust of legal scholars and policymakers. Without exception, the enthusiasm for differential privacy stems from misinformed understanding of how the standard works. This Part also explores instances where differential privacy will likely work well and where it will likely not.

---

100. *See, e.g.*, Chin & Klinefelter, *supra* note 11, at 1452–53; Ohm, *supra* note 10, at 1756; Wu, *supra* note 9, at 1139–40; Felten, *supra* note 12.

101. Cynthia Dwork et al., *Differentially Private Marginals Release with Mutual Consistency and Error Independent of Sample Size*, EUROSTAT WORK SESSION ON STAT. DATA CONFIDENTIALITY 193, 198 (2007).

102. *See* Greengard, *supra* note 2, at 17; Chin & Klinefelter, *supra* note 11, at 1452–55.

103. *See* Chin & Klinefelter, *supra* note 11, at 1422–23. Chin and Klinefelter describe an investigation that they conducted to assess the security practices of Facebook. *Id.* at 1432–45. Based on their analysis, the authors conclude that Facebook is likely using differential privacy, even though Facebook has never indicated that they are. *Id.* at 1422–23. Since the researchers submitted over 30,000 queries, almost any selection of epsilon would have required the noise for each query to dominate the true answer. *See id.* at 1436. Either Facebook is using some other noise-adding mechanism, or the company is implementing differential privacy incorrectly.

A. Misinformed Exuberance

The examples worked through in Part II showed that differential privacy has serious practical limitations. Somehow these problems have escaped the notice of many scholars and journalists, even when the drawbacks are right under their noses.

Consider this excerpt from a *Scientific American* article:

> Suppose the true answer to [a query] is 157. The differentially private algorithm will "add noise" to the true answer; that is, before returning an answer, it will add or subtract from 157 some number, chosen randomly according to a predetermined set of probabilities. Thus, it might return 157, but it also might return 153, 159 or even 292. The person who asked the question knows which probability distribution the algorithm is using, so she has a rough idea of how much the true answer has likely been distorted (otherwise the answer the algorithm spat out would be completely useless to her). However, she doesn't know which random number the algorithm actually added. [104]

This is a typical explanation and endorsement of differential privacy and it makes an equally typical mistake. The author starts with an assumption that contorts the rest of her analysis. The key here is that the reader already knows what the true answer is—157. It is only if the reader already knows the answer that a response like "159 or even 292" can seem useful. But how would the hypothetical researcher, who must operate in ignorance of the true answer, react to a response of "159 or even 292?"

Now consider how the query response in this hypothetical could be meaningful. First, the response might be useful if the selected ε is large, so that the magnitude of the noise is very likely to be small. But the author says the response could very well be 292. If the noise added spans a range of 150, ε in this case cannot be small. We can rule out this possibility.

The second possibility is that a span of 150 might still be small relative to the sort of numbers the researcher was expecting to observe. For example, if the questioner had asked a database containing information on the entire US population to return the number of people who live in particular town in order to understand whether the town is big or small, then a response within 150 of the true value sheds some light. As we have said before, count queries that happen to have very large values are suitable for differential privacy techniques.[105] However, these are unusual conditions. For most researchers, an answer that is likely to be 150 away from the true answer, and that allows them only to conclude things like "this is large-ish" or "this is probably small" will not be good enough. After

104. Klarreich, *supra* note 7.
105. *See supra* Part II.A.

all, from the perspective of a researcher who does not know the true answer, a query response of "292" with a margin of error in excess of 150 would have to consider that the true answer *might* be 442, and that is quite far off from the true answer, which we know to be 157.

The *Scientific American* journalist assumed that the questioner already knew the true answer, or, at least, has a good sense of its ballpark.[106] The experience of a researcher who already knows the answer makes a lousy gauge for the utility of a query system. Instead, we should be concerned about the researchers who potentially do not know what the approximate true answer is. After all, if the researcher knew the approximate answer, he would have little reason to use a query system that adds noise. *Scientific American* thus relays some of the misplaced confidence of the developers of differential privacy.

We take our next example from a Microsoft whitepaper titled *Differential Privacy for Everyone*.[107]

A researcher wants to test whether a particular disease is more likely to manifest in people who have lived in certain regions. She connects to her hospital's query system that has differential privacy guards in place. The researcher makes a series of queries on the number of patients with the disease who have lived in each of the towns in the suspected region. Suppose that some of the towns have a large number of people with the disease, some towns have no people with the disease, and one town, Smallville, has a single case. If the query system were to report the true answers to the researcher, the patient (Bob) in Smallville may be at risk. For example, if he had very recently moved to the researcher's hometown, and the researcher knows he is from Smallville, she might be able to put together that he has the disease. The Microsoft whitepaper explains:

> To avoid this situation, the [query system] will introduce a random but small level of inaccuracy, or distortion, into the results it serves to the researcher. . . .
>
> . . . .
>
> Thus, the answers reported by the [query system] are accurate enough that they provide valuable information to the researcher, but inaccurate enough that the researcher cannot know if Bob's name is or is not in the database.[108]

The conclusions that Microsoft urges us to draw are speculative, to say the least. There is simply no guarantee that the responses from the query system would lead the researcher to the correct approximate understanding about where the cases of the

106. *See* Klarreich, *supra* note 7.
107. *See* MICROSOFT, *supra* note 4, at 4–5.
108. *Id.* at 5.

disease do and do not come from. Whether the responses are only "slightly larger or smaller" will depend entirely on the data curator's specification of ε and the total number of queries.[109]

For good measure, let us quickly work through the hypothetical selecting a relatively liberal value for ε (that is, a less privacy-protecting choice). Suppose $\varepsilon = ln(3)$, which is approximately 1.0986. Assume also that the curator of the database has determined that a total of 1000 simple count queries can be issued to the database. Allowing a range of queries would require us to add more noise, so this is a realistic lower bound in terms of the distortion of results.

With $\varepsilon = 1.0986$ and $m = 1000$, we must use $\varepsilon_q = (1.0986/1000)$ for each individual query. As with all count queries, the most a single individual can influence a count query is by 1, so $\Delta f_q = 1$.[110]

What happens when the researcher queries the system "For each town located in the suspected regions, what is the number of patients with the disease?" Table 9 reports the likelihood that the noise added to each town's response will be within a particular range.

Table 9—Distribution of Laplace Noise Within Specified Ranges $\varepsilon_q = \ln(3)/1000$, $\Delta f = 1$

Noise Range	Probability
±1	0.00
±5	0.01
±10	0.01
±50	0.05
±100	0.10
±500	0.42
±1000	0.67
±10000	1.00

So, for Smallville, there is a very high chance—16%—that the response will exceed 1000, even though we know the true answer is 1. There is also a very high chance—again, 16%—that the response will be less than −1000.

109. Remember that, because the effect on privacy of queries is cumulative, the noise added to each successive query must increase in order to satisfy differential privacy for any specific overall selection of ε. *See supra* note 59 and accompanying text.

110. The noise will be randomly selected from the distribution generated by the Laplace function Lap $(\Delta f_q / \varepsilon_q)$ = Lap(910.239).

Now consider one of the towns "where there are a significant number of individuals" with the disease. Suppose the number of individuals with the disease is about 100. The response has a 45% chance of having a zero or negative value. Even if the number of individuals with the disease in this town is 1000, the probability of observing a negative value response is greater than 16%. Therefore, it is not obvious at all that a faithful use of differential privacy will provide the researcher with meaningful answers from which she could infer that eight towns had a number of people with the disease, and Smallville had either a small number or 0.

To drive this point home, Table 10 provides just one realization, selected randomly from the Laplace noise distribution, for the eight towns and Smallville.

Table 10—Example of Noise-Added Responses to the Smallville Hypothetical $\varepsilon_q = \ln(3)/1000$, $\Delta f = 1$

Town	True Answer	Noise	Response
1	105	2893.9	2998.9
2	80	−2840.6	−2760.6
3	92	848.6	940.6
4	100	4099.3	4199.3
5	125	2145.4	2270.4
6	103	−1607.8	−1504.8
7	99	−814.6	−715.6
8	85	191.3	276.3
Smallville	1	817.3	818.3

The researcher, who sees only the unshaded last column, would be hard-pressed to say anything about the relative prevalence of the disease in these nine towns. The best the researcher could do is conclude that, knowing the value of ε, the true responses were not large enough to overpower the magnitude of the noise that had to be added to maintain differential privacy. The researcher could conclude that none of the towns had tens of thousands of cases of the disease, but she could not confidently say anything more specific than that.

The only practical application of this sort is in response to queries involving common diseases like the flu that occur in the tens of thousands across the subpopulations of interest. For a rare form of cancer, answers drawn from the differential privacy parameters we set will be useless, or worse than useless.[111]

111. Astute readers may notice that the random realization reported in Table 10 is very similar to the output that our fictional internist was confronting in the Introduction. *See supra* note 3 and accompanying text. Indeed, we took the same error drawn here and added it to our

The curator could try to set the parameters differently from ours in order to squeeze some more utility out of the system. The curator could, for example, decide that the system will only respond to a small number of queries so that the ε for each query could be larger. But by reducing the number of queries, the curator reduces the overall value of the query system.[112]

The Microsoft authors' reassurance that "the answers reported by the DP guard are accurate enough that they provide valuable information to the researcher" is thoroughly unwarranted. Reassurances of this sort mislead lay audiences into the optimistic impression that differential privacy preserves data utility better than it does.

By working with examples where they already know the true answer, the proponents of differential privacy have given the impression that the standard is more useful and viable than it really is. Erica Klarreich, the author of the *Scientific American* article, advances the following illustration:

> To see what kind of distribution will ensure differential privacy, imagine that a prying questioner is trying to find out whether I am in a database. He asks, "How many people named Erica Klarreich are in the database?" Let's say he gets an answer of 100. Because Erica Klarreich is such a rare name, the questioner knows that the true answer is almost certainly either 0 or 1, leaving two possibilities:
>
> (a) The answer is 0 and the algorithm added 100 in noise; or
>
> (b) The answer is 1 and the algorithm added 99 in noise.
>
> To preserve my privacy, the probability of picking 99 or 100 must be almost exactly the same; then the questioner will be unable to distinguish meaningfully between the two possibilities.[113]

The assumption that "the questioner knows that the true answer is almost certainly either 0 or 1" turns out to be critical to understanding whether differential privacy is striking the right balance between privacy and utility. We might be satisfied that this intrusive data user must ignore the response to his query because, in the trade-off between his curiosity and Erica Klarreich's privacy, the better interest prevailed.

---

equally fictional "true" responses, which was 20 for each year. Thus, as it turns out, this internist would have had little to worry about if she had known the truth—that seeing a few cases over the course of several weeks is par for the course. Since the internist *did not* know the true values, though, she would have had little reason to feel comforted or alarmed by the responses that she received.

112. There are also some situations in which restricting the database to a small number of queries in order to reduce the magnitude of the noise can produce disclosures. For an example, see Cormode, *supra* note 2, at 1254. These disclosures are not, technically, within Dwork's definition of "disclosure" motivating her differential privacy solutions.

113. Klarreich, *supra* note 7.

But what if the questioner does not know the true answer must be 0 or 1? Instead of "How many people named Erica Klarreich are in the database?" what if the query was "How many people died of postoperative infections last month at this hospital?" Now, when the user receives the response "100," he will either naively assume that the hospital must have terrible sanitary conditions, or, if he is a sophisticated user, he would know to ignore the results since the probability distribution of the noise is in the order of ± 100.

Thus, although we changed nothing about the differential privacy mechanism (altering only the intent of the data user, who in this case is not malicious), a result of "100" to a query whose true result is 0 or 1 is no longer satisfactory. After all, if the true answer is 0, we would not want the data user to worry about the conditions of the hospital. But if the true result were close to 100, we *would* want the researcher to worry. If a hospital were to create a publicly available query system, it would have to anticipate both types of queries—that is, both the intrusive "how many people named Erica Klarreich" query and the postoperative infections query.

The best way to avoid the absurdities is for data curators to ensure that the magnitude of the noise added to a query is comparable to the true answer. But context-driven addition of noise would violate the basic tenets of differential privacy.[114] To satisfy differential privacy, the noise must be independent, not only of the true answer, but also the size of the database.[115] Legal scholars and policymakers have overlooked this drawback.

B. Willful Blindness to Context

One of differential privacy's strongest and most attractive claims is that it can—and in fact must—be applied without considering the specifics of the queried database.[116] But as we saw with the average income example, the blindness to context has harsh consequences. If databases must protect Bill Gates, George Soros, and other highly unusual individuals, then the curator has only two realistic options: give up on utility, or give up on privacy.

When scholars and journalists provide examples of differential privacy in action, they invariably use tables of counts to show how it works.[117] But statistical research often involves the analysis of numerical data. Our examples show that differential privacy is

114. *See* Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91.
115. *See id.*
116. *See id.*
117. *See* Klarreich, *supra* note 7; Chin & Klinefelter, *supra* note 11, at 1433–35.

unlikely to permit meaningful results to queries for averages and correlations unless the data curator selects a very high ε, but in that case, the curator has abdicated his chance to protect privacy.

The natural desire to avoid absurd results has led some supporters of differential privacy to mischaracterize, possibly even misunderstand, what differential privacy demands and to insist that the characteristics of a database, or the answer to a particular query, has some influence over the noise that is added.[118] For example, Felix Wu describes differential privacy as follows:

> The amount of noise depends on the extent to which the answer to the question changes when any one individual's data changes. Thus, asking about an attribute of a single individual results in a very noisy answer, because the true answer could change completely if that individual's information changed. In this case, the answer given is designed to be so noisy that it is essentially random and meaningless. Asking for an aggregate statistic about a large population, on the other hand, results in an answer with little noise, one which is relatively close to the true answer.[119]

Contrary to Wu's assertion, differential privacy noise is *not* a function of the breadth of the query. Because the noise is based on global sensitivity, for all databases that could possibly exist, the noise added to any particular query response must be the same whether the query involves a single person or a million. When it comes to counts and tabular data, the noise added to a query on a large number of people might be less distorting than noise of the same size added to a query on a small number of subjects. But, with other analyses (like correlation), the distortions will be equally severe no matter the n.[120] Lest there be any doubt, Dwork herself has recently insisted, "Our expected error magnitude is constant, independent of n [the number of data subjects responsive to a query]."[121]

A white paper from Microsoft's differential privacy research team makes a similar error.[122] It states:

> Distortion is introduced into the answers a posteriori. That is, the DP guard gets answers based on pristine data, and then mathematically decides the right amount of distortion that needs to be introduced, based on the type of question that was asked, on the size of the database itself, how much its data changes on a regular basis, etc.[123]

Wu and the authors of the Microsoft paper are unwittingly rewriting how differential privacy works. Wu implies that what matters is the influence that a particular piece of information can have *on the particular query that has been submitted.* This would be a

---

118. *See, e.g.*, Wu, *supra* note 9, at 1138.
119. *Id.*
120. *See supra* Part II.D.
121. Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92.
122. *See* Microsoft, *supra* note 4, at 5.
123. *Id.*

fabulous improvement for preserving the utility of a dataset, but it cannot promise differential privacy because a series of queries could reveal changes in the magnitude of the noise that would reveal information about the underlying values.[124] Thus, the technical literature on differential privacy has consistently maintained that the magnitude of the noise must be independent of the size of the data set, the magnitude of the true answer, and the type of query (except in assessing Δf, which requires an assessment of all possible query responses across the universe of possible datasets).[125]

Finally, Ed Felten, Chief Technologist of the Federal Trade Commission, describes differential privacy as if it curbs the amount of error around a particular response. He uses the following example:

> Let's say [an adversary's] best guess, based on all of the available medical science and statistics about the population generally, is that there is a 2% chance that I have diabetes. Now if we give the [adversary] controlled access to my doctor's database, via a method that guarantees differential privacy at the 0.01% level, then the analyst might be able to adjust his estimate of the odds that I have diabetes–but only by a tiny bit. His best estimate of the odds that I am diabetic, which was originally 2%, might go as low as 1.9998% or as high as 2.0002%. The tiny difference of 0.0002% is too small to worry about.
>
> That's differential privacy.[126]

This is not differential privacy at all. An adversary could query the database for the proportion of patients in the doctor's database who have diabetes. This ratio could significantly improve the adversary's guess for Ed Felten's likelihood of having diabetes. This is especially true if the doctor's practice is large enough so that the noise does not drown out the true response.[127] It is also especially true if Ed Felten's doctor specializes in the treatment of diabetics. So Felten's claim can only be correct if we assume that the proportion of individuals with diabetes in his doctor's practice happens to be 2%, just like the general public.

Felten's example illustrates the sort of willful blindness to context that comes from a threat model orientation. By focusing exclusively on the adversary, Felten fails to see the consequences to legitimate research. In a realistic scenario, the number of patients in

---

124. Kobi Nissim et al., *Smooth Sensitivity and Sampling in Private Data Analysis*, *in* STOC'07 Proceedings of the 39th Annual ACM Symposium on Theory of Computing 75, 78 (David S. Johnson & Uriel Feige eds., 2007).

125. *See* Bhaskar et al., *supra* note 2, at 216 ("The amount of noise introduced in the [differentially private] query-response is . . . [i]ndependent of the actual data entries . . . .").

126. Felten, *supra* note 12.

127. Recall that the differentially private noise is independent from the size of the database so that the reported answer approaches the true answer as the size increases.

the doctor's database is likely to be a few thousand.[128] A query system using $\varepsilon = 0.0001$ would have to add tremendous noise to each response.[129] The answers are unlikely to be anywhere close to the true value—whether the legitimate user queries the doctor's database for a count of the number of patients with diabetes or asks point blank "Does Ed Felten have diabetes?" The consequences to research are an afterthought for the proponents of differential privacy.

The legal scholars and policymakers who endorse differential privacy do so only when (and because) they think it works differently than it really does.[130] Differential privacy eschews a nuanced approach that takes into account the variety of disclosures relatively likely to occur, the underlying data, and the specifics of a particular query. This "one size fits all" solution has exactly the problems that one would expect from a nonnuanced rule. It behaves like Procrustes's bed, cutting off some of the most useful applications of a query system without reflection on the costs.

C. Expansive Definitions of Privacy

Differential privacy is motivated by statistician Tore Dalenius's definition of disclosure, which identifies *any* new revelation that can be facilitated by a research database as a reduction of privacy.[131] As Dalenius well knew, eliminating this type of disclosure is not only impossible, it is not even the right goal.[132] Differential privacy makes no differentiation between the types of auxiliary information that an intruder may or may not have. Because it remains agnostic to these types of considerations, the assumptions about what an attacker might know are unrealistic and too demanding. In order to make differential privacy protections manageable, data curators will be tempted to choose a large value for ε or to relax the standards in some other way. But this will relax the privacy protections in a thoughtless way, divorced from context, and thus runs the risk of exposing a few data subjects to unnecessary risks. Embracing too expansive a definition of disclosure creates the danger that curators will deviate from the standard without assessing which disclosures are important

128. "The average US panel size is about 2,300." Justin Altschuler, MD, David Margolius, MD, Thomas Bodenheimer, MD & Kevin Grumbach, MD, *Estimating a Reasonable Patient Panel Size for Primary Care Physicians with Team-Based Task Delegation*, 10 Annals Fam. Med. 396, 396 (2012).

129. The 1% to 99% range of the noise would be approximately −40,000 to +40,000.

130. *See* Wu, *supra* note 9, at 1137–40; Felten, *supra* note 12.

131. Tore Dalenius, *Towards a Methodology for Statistical Disclosure Control*, 5 STATISTISK TIDSKRIFT 429, 433 (1977).

132. *Id.* at 439–40 ("It may be argued that elimination of disclosure is possible only by elimination of statistics.").

(e.g., an increased chance of inferring that Bob has HIV) and which are not (e.g., a decreased chance of inferring that Bob is not a billionaire).

The expansiveness of differential privacy comes from its anticipation of all databases in the universe. Differential privacy defines privacy breach as the gap in probabilities of observing a particular response, not for the particular database in use, but for *all possible datasets* X and X^* that differ on, at most, one row.[133] This is why we have to consider George Soros's income when we are dealing with the income of the citizens of Booneville.

The rationale for this requirement comes from the fact that we not only have to provide protection for the citizens of Booneville, but we must also prevent the response from revealing that someone is *not* a citizen of Booneville. This is true even if it is generally known that George Soros is not a citizen of Booneville and that Booneville does not tend to attract people with wealth. Thus, what may have looked like an advantage of differential privacy—that it requires no assumptions about what adversaries already know—is actually a stumbling block. It causes differential privacy to obliterate accurate responses with noise. By calibrating to the most extreme case (i.e., George Soros), differential privacy protects everyone, but only at significant cost to research.

This explains why differential privacy seems to work pretty well for some counts of individuals but not so well for other variables. For counts, every person exerts the same level of influence and $\Delta f = 1$ regardless of who is or is not included in the database.[134] But for other variables, such as income, the influence exerted by an outlier is very different than that exerted by nearly every other entry. Attempting to protect George Soros's income information adds so much noise that it overwhelms the information about the income of the average citizen (from Booneville or any other city). Dwork obliquely acknowledges as much when she says, "Our techniques work best – i.e., introduce the least noise – when Δf is small."[135] What is left unsaid is that when Δf is very large, differential privacy simply breaks down.

Comparing two databases that differ in one record from the universe of all databases leads to the popularized claim of differential privacy "that it protects against attackers who know all but one

133. *See* Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91–92.

134. *See id.* at 88–89.

135. Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *supra* note 37, at 7.

record."[136] The negative consequences of this requirement are less well known. Differential privacy provides protection in anticipation of the worst-case scenario, which is admirable, but impractical. We could build every building as if it were Fort Knox—but at what cost?

D. Multiple Queries Multiply the Problems

The effect of differential privacy protections on each query is cumulative.[137] This is one of the least discussed factors in the implementation of differential privacy. Any reasonably sized database—such as that of a healthcare provider—is likely to be queried thousands of times. For databases released by government agencies, such as the Census Bureau, the number of queries could easily reach the millions. This is likely true for large databases held by Facebook, Google, and others.[138] If the curator provides responses to a set of m separate queries with privacy parameter $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$, then the global privacy measure for the database is $\varepsilon = \sum_{q=1}^{m} \varepsilon_q$, and thus the differential privacy risk e^{ε}.[139] That is, the differential privacy standard is the sum of all the query epsilons.[140] If the curator wants to keep the global ε under 10, he would have to set either ε_q (the ε for each query) or m (the number of queries) to be quite small. In either case, this severely limits the usefulness of the database. Neither is desirable.

A majority of statistical analyses, such as hypothesis testing, relies on at least the mean and variance—or in the case of multiple variables, the means and the correlations. When every quantity is a "noise-added" response, the effects of large noise-addition can lead to meaningless, or even dangerous, conclusions.

---

136. Daniel Kifer & Ashwin Machanavajjhala, *No Free Lunch in Data Privacy*, *in* SIGMOD '11 Proceedings of 2011 ACM SIGMOD International Conference on Management of Data 193, 193 (2011).

137. *See* Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92.

138. *See* Olanoff, *supra* note 65.

139. *See* Dwork & Smith, *supra* note 14, at 137; Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91.

140. *See* Dwork & Smith, *supra* note 14, at 137; Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92. Returning to Chin and Klinefelter's analysis of responses to 30,000+ different Facebook queries, Chin and Klinefelter conclude that Facebook is likely using a rounding function and a noise addition mechanism that is consistent with $\varepsilon_q = 0.181$ for each query. Chin & Klinefelter, *supra* note 11, at 1433–40. For the set of 30,000+ queries as a whole, this would imply that $\varepsilon = (0.181 \times 30000) = 5430$ which translates into a privacy risk ratio of e^{5430} which is so large that, for all practical purposes, it might as well be infinite. Whether the mistake is Chin and Klinefelter's (for misidentifying differential privacy) or Facebook's (for misapplying it), it shows a frequent, critical failure to understand that the response to every query contributes to the adversary's ability to compromise the privacy of an individual, resulting in wildly overstated descriptions of the privacy offered by differential privacy mechanisms.

E. At the Same Time, Limited Definitions of Privacy

Differential privacy ensures that an individual's inclusion or exclusion from the dataset does not change the probability of receiving a particular query response by too much, but meeting this standard does not necessarily guarantee privacy in the conventional sense.

First, differential privacy leaves the designation of ε to the discretion of the data curator.[141] If the curator is committed both to differential privacy and to maintaining the utility of the data query system, he will be tempted to select a large ε and to allow a large number of queries. If the curator selects a large ε, the standard will be so relaxed that the benefits of differential privacy are wasted. For example, suppose the curator selects $\varepsilon = 10$. 10 sounds like a reasonable enough number, but the privacy standard is actually e^{ε}. So when $\varepsilon = 10$, the ratio of probabilities for a result with and without the inclusion of an individual can be over 22,000. The ratio just need be less than e^{10} (about 22,026.3).[142] With probabilities this different, the curator would have more luck protecting the privacy of the data subjects by adding random noise selected within some context-appropriate bounded range. If the ε is large, the protections offered are hardly worth the effort. The nature of exponents is such that small differences in ε cause very large differences in privacy protection. Table 11 shows the powers of e.

Table 11—Differential Privacy Standards (Ratio of Probabilities) for Varying Selections of ε

ε	e^{ε}			ε	e^{ε}
0.01	1.01			ln(3)	3.00
0.05	1.05			2	7.39
0.10	1.11			5	148.41
0.25	1.28			10	22,026.47
0.50	1.65			25	7.2×10^{10}
ln(2)	2.00			50	5.18×10^{21}
1.00	2.72			100	2.68×10^{43}

Let us work through a quick example of what happens when the curator decides to answer one thousand queries from the

141. *See* Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 88.

142. Even seasoned researchers make the mistake of setting unreasonably high values for ε. For instance, Anne-Sophie Charest sets $\varepsilon =$ at 250, and David McClure and Jerome Reiter set $\varepsilon =$ at 1000, which offers no guarantee of privacy whatsoever. *See* Anne-Sophie Charest, *How Can We Analyze Differentially-Private Synthetic Datasets?*, 2 J. PRIVACY & CONFIDENTIALITY 21, 27 (2010); David McClure & Jerome P. Reiter, *Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data*, 5 TRANSACTIONS ON DATA PRIVACY 535, 536 (2012).

Booneville City database (which may contain, in addition to income, a lot of other information about the citizens of Booneville). For a single query, we observe that the probability of observing a response within ±$1 million is approximately 3% (and 97% of the time it was higher than this range). To be equitable, we assume that every query will be answered with the same level of privacy (assuring both equally accurate responses to all queries and equal privacy for all citizens) resulting in $\varepsilon_q = (\varepsilon/1000)$. This means that the noise added would increase *thousandfold.*[143] With one thousand queries, the observations for average income over a small town would be laughably wrong. The query system would provide responses within *$1billion* of the true answer about 3% of the time. The rest of the time (the remaining 97%), the response will be greater than $1 billion or less than negative $1 billion.

Dwork occasionally underplays the importance of the selection of ε to guard against potential privacy-invading uses. She states "if the [differentially private] database were to be consulted by an insurance provider before deciding whether or not to insure a given individual, then the presence or absence of any individual's data in the database will not significantly affect his or her chance of receiving coverage."[144] But with a high enough ε, an insurance adjustor could take advantage of the lax standard. For example, suppose the adjustor asks, "Does Jeff Jones have a congenital heart disease?" and ε is set to $ln(2)$. This means that the ratio of probabilities that the database will give a particular response equals 2. Thus, if Jeff Jones were to have the disease, it is twice as likely to observe a response that he has the diseases compared to the response that he does not have the disease.[145] So when they receive a positive response, the insurance company may want to play the odds and decline coverage.

The effects are worse for clusters of individuals. Consider an insurance company employee who issues the query, "How many individuals in the Jones family of 5 have a congenital heart disease?" Assuming one or more of the individuals in this family does have the congenital heart disease, the probability of a response indicating that

143. One of the interesting aspects of the Laplace distribution is that the noise for m queries is a direct multiple of the noise for one query. The Laplace inverse cumulative distribution function with mean zero is written as: $-b \times Sign(p - 0.5) \times ln(1 - 2|p - 0.5|)$ where b is the shape parameter of the Laplace distribution and p is a random number between 0 and 1. When a single query is answered, $b = \Delta f/\varepsilon$ and when m queries are answered $b' = \Delta f/(\varepsilon/m) = m(\Delta f/\varepsilon) = mb$. For a given random number p, the noise using b' is m times the noise generated using b.

144. Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91 (emphasis omitted).

145. *See id.* at 91–92.

one or more individuals in this family has the disease is 32 times (2^5) more likely than a negative response because differential privacy ensures only that each marginal individual contribute no more than a doubling of the probability. For five individuals in a row, the ratio would double five times. Now, the insurance adjustor is very likely to decline coverage for the Jones family since the chance that all of them *don not* have heart disease may be a paltry 1/33.[146]

F. Difficult Application

Because differential privacy techniques are agnostic to the specific underlying database, one might get the impression that they are easy to implement. This is not the case.

In order to create the appropriate Laplace noise distribution, the data curator must identify and assess the global sensitivity (Δf) for every type of allowable query.[147] For some statistics, such as counts, sums, and mean, the analysis is straightforward. For most tabular data, $\Delta f = 1$.[148] Sums and means require the curator to know the largest values over the entire world's population for each variable, but as long as they have access to some reliable descriptive statistics[149], this is usually not too hard.

For analyses involving more complicated statistics, determining global sensitivity is not an easy task. Consider the illustration in which a user queries a database for the average income of residents in Booneville, Kentucky. In order to compute Δf, the data steward will have to guess the income of the world's highest-paid person. Error has serious consequences: under-specifying Δf would mean that differential privacy is not actually satisfied, but over-specifying Δf will further degrade the quality of the output. Statistical analysis often involves estimates of important statistical

146. Graham Cormode also provides an interesting example of a disclosure that can be made while satisfying differential privacy, but which is avoidable with more traditional, context-driven privacy measures. Cormode, *supra* note 2, at 1256–57.

147. *See* Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92.

148. Even tabular data has the potential to cause confusion. Klarreich, author of the *Scientific American* article, provides an illustration of a type of disclosure that occurs with genotype frequencies. Klarreich, *supra* note 7. Unfortunately, in this situation, it would not be possible to maintain the privacy parameter for each cell and the overall database at ε. The data involves frequencies of thousands of *different* single nucleotide polymorphisms (SNPs) and every individual is represented in every SNP frequency. *See id.* The addition/deletion of one record will modify every one of the SNP frequencies. To see an attack taking advantage of these circumstances, see Daniel I. Jacobs et al., *Leveraging Ethnic Group Incidence Variation to Investigate Genetic Susceptibility to Glioma: A Novel Candidate SNP Approach*, 3 FRONTIERS IN GENETICS 203, 203 (2012).

149. Hopefully the curator's source for learning the global range does not employ differential privacy.

relationships between numerical variables such as variance, regression coefficients, coefficient of determination, or eigen-values. For these types of queries, determining global sensitivity will be very challenging. Correctly choosing global sensitivity has drastic consequences to utility—as we saw with the correlation example in Part II.

Considering all of these limitations together, we must circumscribe the practical applications for pure differential privacy to the situations in which count queries have true answers that are very large. Unless we alter the core purposes and definitions of differential privacy, statisticians and policymakers should ignore the hype.

CONCLUSIONS

Differential privacy faces a hard choice. It must either recede into the ash heap of theory or surrender its claim to uniqueness and supremacy. In its pure form, differential privacy has no chance of broad application. However, recent research by its proponents shows a willingness to relax the differential privacy standard in order to complex queries. Two such relaxations are often used.

The first, proposed by Dwork herself, requires that the probability of seeing a response with a particular subject remain within some factor of the probability of the same response without that subject *plus* some extra allowance.[150] The problem with this modification is that there is no upper bound on the actual privacy afforded by this standard.[151] In some situations, this allowance may be appropriate, but it would require the judgment of a privacy expert based on context—the very thing differential privacy had sought to avoid.

Ashwin Machanavajjhala developed another alternative for the US Census Bureau's On the Map application.[152] This relaxation of differential privacy allows curators to satisfy a modified differential privacy standard while usually meeting strict differential privacy. For some predesignated percentage of responses, the differential privacy

150. *See* Dwork & Smith, *supra* note 14, at 139. Mathematically, the relationship looks like this:

$P(R = r|X) \leq e^{\varepsilon} \times P(R = r|X^*) + \delta$ where δ is small.

151. The extent to which actual probability ratio is different from the ratio that includes or excludes a data subject is bounded by the $e^{\varepsilon} + \frac{\delta}{P(R=r|X^*)}$, but when $P(R = r|X^*)$ is very small (say 0.00001) and $\delta = 0.01$, the privacy ratio can exceed differential privacy standards by 1000. Even though δ is small, the risk of disclosure can be very large.

152. Ashwin Machanavajjhala et al., *Privacy: Theory Meets Practice on the Map*, *in* ICDE '08 PROCEEDINGS OF THE 2008 IEEE 24TH INTERNATIONAL CONFERENCE ON DATA ENGINEERING 277, 283 (2008).

standard can be broken.[153] This relaxation also undermines the promise of privacy.[154] In the situations where differential privacy is not satisfied, there is no upper bound on the risk of disclosing sensitive information to a malicious user. However, this may be fine if the curator crafts the deviations in a thoughtful way. Nonetheless, the data curator would need to resort to judgment and context.

This progression by the differential privacy researchers to a relaxed form is odd, given their view that historical definitions of privacy in the statistical literature lack rigor. The differential privacy community roundly dismisses traditional mechanisms for not offering strong privacy guarantees,[155] but the old methods will often satisfy the proposed relaxed forms of differential privacy as On the Map clearly illustrates.[156]

As differential privacy experts grapple with the messy problems of creating a system that gives researchers meaningful responses, while also providing meaningful disclosure prevention—albeit not differential privacy—they have come back to earth and rejoined the rest of the disclosure risk researchers who toil with the tension between utility and privacy.[157] In its strictest form, differential privacy is a farce. In its most relaxed form, it is no different, and no better, than other methods.[158]

Legal scholars and policymakers should resist the temptation to see differential privacy as a panacea, and to reject old disclosure prevention methods as inadequate. Adopting differential privacy as a regulatory best practice or mandate would be the end of research as we know it. The answers to basic statistical questions—averages and correlations—would be gibberish, and the standard would be very difficult to apply to regression and other complex analyses.

---

153. *See id.* at 280–81.

154. For example, the authors go on to propose a relaxation of differential privacy that satisfies differential privacy albeit with $\varepsilon = 8.6$, which implies a privacy risk ratio of $e^{8.6} = 5431.66$. *Id.* at 284. This implies that, based on the responses (or in this case released data), we can conclude the presence of an individual has probability that can be 5431.66 times higher than the absence of an individual.

155. *See* Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *supra* note 37, at 1 (criticizing disclosure prevention mechanisms for being syntactic and ad hoc).

156. *See* Machanavajjhala et al., *supra* note 152, at 277.

157. *See, e.g.*, Bhaskar et al., *supra* note 2, at 216 ("While the form of our guarantee is similar to DP, where the privacy comes from is very different, and is based on: 1) A statistical (generative) model assumption for the database, 2) Restrictions on the kinds of auxiliary information available to the adversary.").

158. For example, differential privacy offers no greater security against Dinur-Nissim "blatant non-privacy" unless the data curator strictly limits the number of queries that can be issued to the system. *Cf.* Dinur & Nissim, *supra* note 23, at 203–04, 206. Other noise-adding approaches, too, can avoid the Dinur-Nissim results by limiting the number of queries. *See supra* note 28 and accompanying text.

Differential privacy would also forbid public microdata releases—a valuable public information resource.[159] Lest we end up in a land with a negative population of 30 foot-tall people earning an average income of $23.8 million per year, the legal and policy community must curb its enthusiasm for this trendy theory.

159. *See* Barbara J. Evans, *Much Ado About Data Ownership*, 25 HARV. J.L. & TECH. 69, 76, 94 (2011) (discussing the value of compiling patient metadata for research).