DEFINING PRIVACY AND UTILITY IN DATA SETS

FELIX T. WU*

Is it possible to release useful data while preserving the privacy of the individuals whose information is in the database? This question has been the subject of considerable controversy, particularly in the wake of well-publicized instances in which researchers showed how to re-identify individuals in supposedly anonymous data. Some have argued that privacy and utility are fundamentally incompatible, while others have suggested that simple steps can be taken to achieve both simultaneously. Both sides have looked to the computer science literature for support.

What the existing debate has overlooked, however, is that the relationship between privacy and utility depends crucially on what one means by "privacy" and what one means by "utility." Apparently contradictory results in the computer science literature can be explained by the use of different definitions to formalize these concepts. Without sufficient attention to these definitional issues, it is all too easy to overgeneralize the technical results. More importantly, there are nuances to how definitions of "privacy" and "utility" can differ from each other, nuances that matter for why a definition that is appropriate in one context may not be appropriate in another. Analyzing these nuances exposes the policy choices inherent in the choice of one definition over another and thereby elucidates decisions about whether and how to regulate data privacy across varying social contexts.

---

* Associate Professor, Benjamin N. Cardozo School of Law. Thanks to Deven Desai, Cynthia Dwork, Ed Felten, Joe Lorenzo Hall, Helen Nissenbaum, Paul Ohm, Boris Segalis, Kathy Strandburg, Peter Swire, Salil Vadhan, Jane Yakowitz, and participants at the 2011 Privacy Law Scholars Conference, the 2012 Works-In-Progress in IP Conference, the 2012 Technology Policy Research Conference, the New York City KnowledgeNet meeting of the International Association of Privacy Professionals, the NYU Privacy Research Group, the Washington, D.C. Privacy Working Group, and the Harvard Center for Research on Computation and Society seminar for helpful comments and discussions.

INTRODUCTION

The movie rental company Netflix built its business in part on its ability to recommend movies to its customers based on their past rentals and ratings. In 2006, Netflix set out to improve its movie recommendation system by launching a contest.[1] The company challenged researchers throughout the world to devise a recommendation system that could beat its existing one by at least 10 percent, and it offered one million dollars to the team that could exceed that benchmark by the widest margin.[2] "Anyone, anywhere" could register to participate.[3] Participants were given access to a "training data set consist[ing] of more than 100 million ratings from over 480 thousand randomly-chosen, anonymous customers on nearly 18

1. *See The Netflix Prize Rules*, NETFLIX PRIZE, http://www.netflixprize.com/rules (last visited Feb. 16, 2013).
2. *Id.*
3. *Id.*

thousand movie titles."[4] Researchers could use this data to train the recommendation systems they designed, which were then tested on a set of additional movies rated by some of these same customers, to see how well a new system predicted the customers' ratings. More than forty thousand teams registered for the contest, and over five thousand teams submitted results.[5] Three years later, a team of researchers from AT&T Research and elsewhere succeeded in winning the grand prize.[6] Netflix announced plans for a successor contest, which would use a data set that included customer demographic information, such as "information about renters' ages, gender, ZIP codes, genre ratings[,] and previously chosen movies."[7]

Meanwhile, a team of researchers from the University of Texas registered for the contest with a different goal in mind. Rather than trying to predict the movie preferences of the customers in the data set, these researchers attacked the problem of trying to figure out who these customers were.[8] Netflix, having promised not to disclose its customers' private information[9] and perhaps recognizing that it might be subject to the Video Privacy Protection Act,[10] had taken steps to "protect customer privacy" by removing "all personal information identifying individual customers" in the data set and replacing all customer identification numbers with "randomly-assigned ids."[11] Moreover, to further "prevent certain inferences [from] being drawn about the Netflix customer base," Netflix had also "deliberately perturbed" the

4. *Id.*

5. *See Netflix Prize Leaderboard*, NETFLIX PRIZE, http://www.netflixprize.com/leaderboard (last visited Feb. 16, 2013); *see also BellKor's Pragmatic Chaos*, AT&T LABS RESEARCH, http://www2.research.att.com/~volinsky/netflix/bpc.html (last visited Feb. 16, 2013) (describing the members of the winning team, BellKor's Pragmatic Chaos).

6. *See* Steve Lohr, *Netflix Awards $1 Million Prize and Starts a New Contest*, N.Y. TIMES (Sept. 21, 2009), http://bits.blogs.nytimes.com/2009/09/21/netflix-awards-1-million-prize-and-starts-a-new-contest/.

7. *Id.*

8. *See* Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Datasets*, 29 PROC. IEEE SYMPOSIUM ON SECURITY & PRIVACY 111, 111–12 (2008).

9. *See* Complaint at 7, Doe v. Netflix, Inc., No. 09-cv-0593 (N.D. Cal. Dec. 17, 2009) ("Except as otherwise disclosed to you, we will not sell, rent or disclose your personal information to third parties without notifying you of our intent to share the personal information in advance and giving you an opportunity to prevent your personal information from being shared.") (quoting Netflix's then-current Privacy Policy).

10. *See* 18 U.S.C. § 2710 (2012).

11. *The Netflix Prize Rules*, *supra* note 1.

data set by "deleting ratings, inserting alternative ratings and dates, and modifying rating dates."[12] The Texas researchers showed, however, that despite the modifications made to the released data, a relatively small amount of information about an individual's movie rentals and preferences was enough to single out that person's complete record in the data set.[13] In other words, someone who knew a little about a particular person's movie watching habits, such as might be revealed in an informal gathering or at the office, could use that information to determine the rest of that person's movie watching history, perhaps including movies that the person did not want others to know that he or she watched.[14] Narayanan and Shmatikov also showed that sometimes the necessary initial information could be gleaned from publicly available sources, such as ratings on the Internet Movie Database.[15]

After Narayanan and Shmatikov published their results, a class action lawsuit was filed against Netflix, in which the plaintiff class alleged that the disclosure of the Netflix Prize data set was a disclosure of "sensitive and personal identifying consumer information."[16] The lawsuit later settled on undisclosed terms.[17] As part of the settlement, Netflix agreed to scrap the successor contest,[18] and it removed the original data set from the research repository to which it had previously given the information.[19]

What is the lesson of the Netflix Prize story? Does it herald a new era in the science of data analysis, in which data release inevitably leads to tremendous privacy loss? Or is it an outlier event that should be dismissed as inconsequential to law and policy going forward?

---

12. *Id.*

13. *See* Narayanan & Shmatikov, *supra* note 8, at 121 ("[V]ery little auxiliary information is needed [to] de-anonymize an average subscriber record from the Netflix Prize dataset. With eight movie ratings (of which two may be completely wrong) and dates that may have a fourteen-day error, ninety-nine percent of records can be uniquely identified in the dataset.").

14. *See id.* at 122.

15. *See id.* at 122–23.

16. Complaint, Doe v. Netflix, *supra* note 9, at 2; *see also* Ryan Singel, *Netflix Spilled Your* Brokeback Mountain *Secret, Lawsuit Claims*, WIRED (Dec. 17, 2009), http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit/.

17. *See* Ryan Singel, *NetFlix Cancels Recommendation Contest After Privacy Lawsuit*, WIRED (Mar. 12, 2010), http://www.wired.com/threatlevel/2010/03/netflix-cancels-contest/.

18. *See id.*

19. *See Note from Donor Regarding Netflix Data*, UCI MACHINE LEARNING REPOSITORY (Mar. 1, 2010), http://archive.ics.uci.edu/ml/noteNetflix.txt.

Neither of those extreme answers is correct. Rather, the narrow lesson of the story is that releasing data that is useful in a particular way turns out to be less private than we thought. The broader lesson to be learned, of which the Netflix Prize story is only a part, is that there are many different senses in which data can be useful and in which a data release can be private. In order to set appropriate data policy, we must recognize these differences, so that we can explicitly choose among the different conceptions.

When Netflix released its data set, it thought that it could serve two goals simultaneously: protecting the privacy of its subscribers, while enabling valuable research into the design of recommendation systems. In other words, Netflix was trying to release data that was both private and useful. These twin goals of privacy and utility can be in tension with each other. Information is useful exactly when it allows others to have knowledge that they would not otherwise have and to make inferences that they would not otherwise be able to make. The goal of information privacy, meanwhile, is precisely to prevent others from acquiring particular information or from being able to make particular inferences.[20]

There is nothing inherently contradictory, however, about hiding one piece of information while revealing another, so long as the information we want to hide is different from the information we want to disclose. In the Netflix case, the contest participants were aimed at one goal, predicting movie preferences, while Narayanan and Shmatikov were aimed at a different one, uncovering customer identities. The promise of anonymization is that, by removing "personally identifiable information" and otherwise manipulating the data, the released information can be both useful for legitimate purposes and private.[21]

In the Netflix example, as well as in other prominent

20. At least, that is the relevant goal for purposes of the problems described in this article. In general, the word "privacy" has been used to describe a wide variety of goals that may not have a single distinguishing feature. *See generally* DANIEL J. SOLOVE, UNDERSTANDING PRIVACY (2008).

21. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1707–11 (2010). Different laws or commentators refer alternatively to either "anonymized" and "de-anonymized" data or "identified," "de-identified," and "re-identified" data. *See, e.g.*, *id.* at 1703. Although in fact different uses of these terms may refer to different concepts, see *infra* Part III.A, the terminology does not track these differences, and this article also uses both sets of terminology interchangeably.

examples,[22] anonymization seems not to have worked as intended, and researchers have been able to "de-anonymize" the data, thereby learning the information of particular individuals from the released data. These examples of de-anonymization have led some to argue that privacy and utility are fundamentally incompatible with each other and that supposedly anonymized data is never in fact anonymous.[23] On this view, the law should never distinguish between "personally identifiable" information and "anonymized" or "de-identified" information, and regulators should be wary of any large-scale, public data releases.[24]

Others, though, have characterized the existing examples of de-anonymization as outliers, and have argued that straightforward techniques suffice to protect against any real risks of re-identification, while still making useful research possible.[25] These commentators have argued that identifying a category of de-identified information that can be freely shared is still the right approach and that too much reluctance to release de-identified data will stunt important research in medicine, public health, and social sciences, with little benefit to privacy interests.[26] More recently, some have argued that what the law needs is a three-tiered system in which the level of data privacy regulation depends on whether the data poses a "substantial," "possible," or "remote" risk of re-identification.[27]

The question of how to define and treat "de-identified" data, as opposed to "personally identifiable" data, is important and pervasive in privacy law.[28] The scope of a wide range of privacy laws depends on whether particular information is "individually identifiable,"[29] "personally identifiable,"[30] or

22. *See* Michael Barbaro & Tom Zeller, Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES, Aug. 9, 2006, at A1; Latanya Sweeney, *k-Anonymity: A Model for Protecting Privacy*, 10 INT'L J. UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYSTEMS 557, 558–59 (2002).

23. *See* Ohm, *supra* note 21, at 1705–06.

24. *See id.* at 1765–67.

25. *See* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1 (2011).

26. *See id.* at 4.

27. Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814, 1877–78 (2011).

28. *See id.* at 1827 (describing the concept of personally identifiable information as having "become the central device for determining the scope of privacy laws").

29. For example, the HIPAA Privacy Rule applies to "protected health information," defined as "individually identifiable" health information. 45 C.F.R.

"personal."[31] Much hinges therefore on whether any such concept is a sensible way of defining the scope of privacy laws, and if so, what that concept should be.

Unsurprisingly then, concerns about whether de-identification is ever effective have begun to manifest themselves in a variety of legal contexts. Uncertainty over whether identifiable data can be distinguished from de-identified data underlies several of the questions posed in a recent advanced notice of proposed rulemaking about possible changes to the Common Rule, which governs human subjects protection in federally funded research.[32] Arguments about the ineffectiveness of de-identification also formed the core of several amicus briefs filed before the Supreme Court in *Sorrell v. IMS Health*, a case involving the disclosure and use of de-identified prescription records.[33] The argument has been used

§ 160.103 (2013). Similarly, the Federal Policy for the Protection of Human Subjects (the "Common Rule") states that "[p]rivate information must be individually identifiable . . . in order for obtaining the information to constitute research involving human subjects." *Id.* § 46.102 (emphasis omitted); *see also Federal Policy for the Protection of Human Subjects ("Common Rule")*, U.S. DEP'T OF HEALTH & HUMAN SERVICES, http://www.hhs.gov/ohrp/humansubjects/commonrule/index.html (last visited Feb. 16, 2013) (noting that the Common Rule is "codified in separate regulations by fifteen Federal departments and agencies" and that each codification is "identical to [that] of the HHS codification at 45 CFR part 46, subpart A").

30. For example, the Video Privacy Protection Act prohibits the knowing disclosure of "personally identifiable" video rental information. 18 U.S.C. § 2710(b)(1) (2006).

31. For example, the Massachusetts data breach notification statute applies when "the personal information of [a Massachusetts] resident was acquired or used by an unauthorized person or used for an unauthorized purpose." MASS. GEN. LAWS ch. 93H, § 3 (2012). Similarly, the E.U. Data Protection Directive applies to the "processing of personal data." Directive 95/46/EC, on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, art. 3, 1995 O.J. (L 281) 31, 39.

32. *See* Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, 76 Fed. Reg. 44512, 44524–26 (July 26, 2011) ("[W]e recognize that there is an increasing belief that what constitutes 'identifiable' and 'deidentified' data is fluid; rapidly evolving advances in technology coupled with the increasing volume of data readily available may soon allow identification of an individual from data that is currently considered deidentified.").

33. *See* 131 S. Ct. 2653 (2011); Brief of Amicus Curiae Electronic Frontier Foundation in Support of Petitioners at 12, *Sorrell*, 131 S. Ct. 2653 (No. 10-779) ("The PI Data at issue in this case presents grave re-identification issues."); Brief of Amici Curiae Electronic Privacy Information Center (EPIC) et al. in Support of the Petitioners at 24, *Sorrell*, 131 S. Ct. 2653 (No. 10-779) ("Patient Records are At Risk of Being Reidentified"); Brief for the Vermont Medical Society et al. as Amici Curiae Supporting Petitioners at 23, *Sorrell*, 131 S. Ct. 2653 (No. 10-779) ("Patient De-Identification of Prescription Records Does Not Effectively Protect

in the context of consumer class actions, claiming that the release of de-identified data breached a promise not to disclose personally identifiable information.[34] A recent consumer privacy report from the Federal Trade Commission (FTC) contains an extensive discussion of identifiability and its effect on the scope of the framework developed in that document.[35]

This legal and policy debate has taken place in the shadow of a computer science literature analyzing both techniques to protect privacy in databases and techniques to circumvent those privacy protections. Legal commentators have invariably cited the science in order to justify their conclusions, even while offering very different policy perspectives.[36] A closer look at the computer science, however, reveals that several aspects of that literature have been either misinterpreted, or at least overread, by legal scholars.[37] There is little support for the strongly pessimistic view that, as a technical matter, "any data that is even minutely useful can never be perfectly anonymous, and small gains in utility result in greater losses for privacy."[38] On the other hand, we should not be too sure that it would be straightforward to "create a low-risk public dataset" that maintains all of the research benefits of the original dataset with minimal privacy risk.[39] Nor should we assume that "metrics for assessing the risk of identifiability of information" will add substantially to the precision of such a risk assessment.[40]

More fundamentally, disagreements over the meaning of the science and the resulting policy prescriptions are rooted in disagreements over the very concepts of "privacy" and "utility" themselves. The apparently competing claims that "as the

Patient Privacy"); *cf.* Brief for Khaled El Emam and Jane Yakowitz as Amici Curiae for Respondents at 2, *Sorrell*, 131 S. Ct. 2653 (No. 10-779) ("Petitioner Amici Briefs overstate the risk of re-identification of the de-identified patient data in this case.").

34. *See, e.g.*, Steinberg v. CVS Caremark Corp., No. 11–2428, 2012 WL 507807 (E.D. Pa. Feb. 16, 2012); Complaint, Doe v. Netflix, *supra* note 9.

35. *See* FEDERAL TRADE COMMISSION, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE 18–22 (2012).

36. *See* Ohm, *supra* note 21, at 1751–58 (explaining why "technology cannot save the day, and regulation must play a role"); Yakowitz, *supra* note 25, at 23–35 (describing "five myths about re-identification risk"); *see also* Schwartz & Solove, *supra* note 27, at 1879 (asserting that "practical tools also exist for assessing the risk of identification").

37. *See infra* Parts I–II.

38. Ohm, *supra* note 21, at 1755.

39. Yakowitz, *supra* note 25, at 54.

40. Schwartz & Solove, *supra* note 27, at 1879.

utility of data increases even a little, the privacy plummets"[41] and that "contemporary privacy risks have little to do with anonymized research data"[42] turn out to be incomparable because the word "privacy" is being used differently in each. One refers to the ability to hide even uncertain information about ourselves from people close to us; the other refers to the ability to prevent strangers from picking out our record in a data set.[43]

Recognizing that there are competing definitions of privacy and utility is only the first step. What policymakers ultimately need is guidance on how to choose among these competing definitions. Accordingly, this Article develops a framework designed to highlight dimensions along which definitions of privacy and utility can vary. By understanding these different dimensions, policymakers will be better able to fit the definitions of privacy and utility to the normative goals of a particular context, better able to find the technical results that apply to the context, and better able to decide whether technical or legal tools will be most effective in achieving the relevant goals.

On the privacy side, the computer science literature provides a good model in framing the issue as one of determining the potential threats to be protected against.[44] Privacy that protects against stronger, more sophisticated, more knowledgeable attackers is a stronger notion of privacy than one that only protects against relatively weaker attackers. Thinking in terms of threats provides the bridge between mathematical or theoretical definitions of privacy and privacy in practice. Defining the relevant threats is also central to understanding how to regard partial, or uncertain, information, such as a 50 percent certainty that a given individual has a particular disease, for example.[45]

If on the privacy side we need to be more specific about what we want to prevent in the wake of a data release, on the utility side we need to be more specific about what we want to make possible. Some types of data processing are more privacy-invading than others.[46] Depending on the context, then, it may

41. Ohm, *supra* note 21, at 1751.
42. Yakowitz, *supra* note 25, at 36.
43. *See infra* Parts I–II.
44. *See infra* Part III.A.
45. *See infra* Part III.B.
46. *See infra* Part III.C.

be important to determine whether the definition of utility needs to encompass particularly complex or particularly individualized data processing. Moreover, it matters a great deal whether we want to allow the broadest possible range of future data uses, or whether it would be acceptable to limit future uses to some pre-defined set of foreseeable uses.[47]

One cannot talk about the success or failure of anonymization in the abstract. Anonymization encompasses a set of technical tools that are effective for some purposes, but not others. What matters is how well those purposes match the law and policy goals society wants to achieve. That is a question of social choice, not mathematics.

Part I below begins by explaining why detractors of anonymization have overstated their case and why the computer science literature does not establish that anonymization inevitably fails. Part II then explains why the flaws of anonymization are nevertheless real and why anonymization should not be seen as a silver bullet. Part III steps back from the debate over anonymization to develop a framework for understanding different conceptions of privacy and utility in data sets, focusing on four key dimensions: (1) defining the relevant threats against which protection is needed; (2) determining how to treat information about individuals that is uncertain; (3) characterizing the legitimate uses of released data; and (4) deciding when to value unpredictable uses. Part IV applies the framework to two specific examples. A brief conclusion follows.

I. WHY WE SHOULDN'T BE TOO PESSIMISTIC ABOUT ANONYMIZATION

In Paul Ohm's leading paper, he argues that privacy law has placed too much faith in the ability of anonymization techniques to ensure privacy.[48] According to Ohm, technologists and regulators alike have embraced the belief "that they could robustly protect people's privacy by making small changes to their data," but this belief, Ohm argues, "is deeply flawed."[49] The flaw is supposedly not just a flaw in the existing techniques, but a flaw in the very idea that technology

47. *See infra* Part III.D.
48. *See* Ohm, *supra* note 21, at 1704.
49. *Id.* at 1706–07.

can be used to balance privacy and utility.[50] Ohm claims that the computer science literature establishes that "any data that is even minutely useful can never be perfectly anonymous, and [that] small gains in utility result in greater losses for privacy."[51]

Ohm's views on the inevitable failure of anonymization have been very influential in recent privacy debates and cases.[52] His article is regularly cited for the proposition that utility and anonymity are fundamentally incompatible.[53] His ideas have also been extensively covered by technology news sites and blogs.[54] Then-FTC Commissioner Pamela Harbour specifically called attention to the article during remarks at an FTC roundtable on privacy, highlighting the possibility that "companies cannot truly deliver and consumers cannot expect anonymization."[55]

A simple thought experiment, however, shows that the truth of Ohm's broadest claims depends on how one conceptualizes privacy and utility. Imagine a (fictitious) master database of all U.S. health records. Suppose a researcher is interested in determining the prevalence of lung cancer in the

50. *See id.* at 1751.

51. *Id.* at 1755.

52. *See, e.g.*, FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: PRELIMINARY FTC STAFF REPORT 38 (2010) (citing Ohm, *supra* note 21); Brief for Petitioners at 37 n.11, Sorrell v. IMS Health Inc., 131 S. Ct. 2653 (2011) (No. 10-779) (same); Brief of Amicus Curiae Electronic Frontier Foundation in Support of Petitioners at 10, *Sorrell*, 131 S. Ct. 2653 (No. 10-779) (same); Brief for the Vermont Medical Society et al. as Amici Curiae Supporting Petitioners at 26, *Sorrell*, 131 S. Ct. 2653 (No. 10-779) (same); *see also* Consolidated Answer to Briefs of Amici Curiae Dwight Aarons et al. at 10, Sander v. State Bar of Cal., 273 P.3d 1113 (Cal. review granted Aug. 25, 2011) (No. S194951) ("*Amici* assert that effective anonymization of records based on information obtained from individuals is impossible . . . . Although they cite a number of authorities for this proposition, they all rely primarily on a single source: a law review article by Paul Ohm entitled *Broken Promises of Privacy* . . . .").

53. *See, e.g.*, JeongGil Ko et al., *Wireless Sensor Networks for Healthcare*, 98 PROC. IEEE 1947, 1957 (2010) ("Data can either be useful or perfectly anonymous, but never both.") (quoting Ohm, *supra* note 21, at 1704).

54. *See, e.g.*, Nate Anderson, *"Anonymized" Data Really Isn't—And Here's Why Not*, ARS TECHNICA (Sept. 8, 2009, 7:25 AM), http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin; Melanie D.G. Kaplan, *Privacy: Reidentification a Growing Risk*, SMARTPLANET (Mar. 28, 2011, 2:00 AM), http://www.smartplanet.com/blog/pure-genius/privacy-reidentification-a-growing-risk/5866; Andrew Nusca, *Your Anonymous Data Is Not So Anonymous*, ZDNET (Mar. 29, 2011, 9:57 AM), http://www.zdnet.com/blog/btl/your-anonymous-data-is-not-so-anonymous/46668.

55. FED. TRADE COMM'N, TRANSCRIPT OF SECOND ROUNDTABLE ON EXPLORING PRIVACY 14–15 (2010).

U.S. population. Is it possible to release data from which this can be calculated, while still preserving the privacy of the individuals in the database? The answer would seem to be yes, since the database administrator can simply release only the number that the researcher is looking for and nothing more.

If that answer is not satisfactory, it must be for one of two reasons. One possibility is that even this single statistic about the prevalence of lung cancer fails to be "perfectly anonymous."[56] Suppose I know the lung cancer status of everyone in the population *except* for the one person I am interested in. Then information about the overall prevalence of lung cancer is precisely the missing link I need to determine the status of the last person.[57] If such a possibility counts as a privacy violation, then the statistic fails to be perfectly private. Moreover, even without any background information, the statistic by itself conveys some information about everyone in the U.S. population. Take a random stranger in the database. If the overall prevalence of lung cancer is one percent, I now "know," with one percent certainty, that this person has lung cancer. If such knowledge violates the random stranger's privacy, then again the statistic fails to be perfectly private. Thus, whether the statistic should be regarded as private depends on how we define "private."

Alternatively, perhaps the statistic fails to be "even minutely useful."[58] In theory, this might be because the calculation of such a statistic falls outside a conception of what it means to conduct research,[59] although this seems unlikely in this particular example. A stronger potential objection here is that a single statistic is too limited to be useful. It answers only a single question and fails to answer the vast number of other questions that a researcher might legitimately ask of the data set.[60] To take that view, however, is again to have a particular

56. *See* Ohm, *supra* note 21, at 1755.

57. This sort of example is precisely what the definition of differential privacy is designed to exclude. *See infra* Part I.B.

58. *See* Ohm, *supra* note 21, at 1755.

59. *See infra* Part III.C.

60. *See infra* Part III.D. It appears that Ohm takes this view. Ohm distinguishes between "release-and-forget anonymization" and the release of "summary statistics," agreeing that the latter can preserve privacy. Ohm, *supra* note 21, at 1715–16. However, the difference between the two is a matter of degree, not of kind. Data that have been subject to enough generalization and suppression eventually become an aggregate statistic. In the example above, if the data administrator suppresses every field except the health condition, and generalizes the health condition to "lung cancer" or "not lung cancer," then the

idea of what it means to be "useful."

Ohm draws his conclusion about the fundamental incompatibility of privacy and utility from the computer science literature.[61] In so doing, he misinterprets important aspects of that literature, both with respect to the impossibility results he cites[62] and with respect to recent research in the area of differential privacy.[63] More importantly, he implicitly adopts the assumptions made in the literature he cites about the nature of privacy and utility, assumptions that are not necessarily warranted across all contexts.

A. Impossibility Results

In support of his claim that privacy and utility inevitably conflict, Ohm relies primarily on a paper by Justin Brickell and Vitaly Shmatikov that purports to "demonstrate that even modest privacy gains require almost complete destruction of the data-mining utility."[64] Despite the broad claims of the Brickell-Shmatikov paper, however, its results are far more modest than Ohm suggests.[65]

Consider the figure that Ohm reproduces in his paper, also reproduced below as Figure 1.[66] As Ohm describes it, for each pair of bars, "the left, black bar represents the privacy of the data, with smaller bars signifying more privacy," while the "right, gray bars represent the utility of the data, with longer

---

resulting data set reveals the prevalence of lung cancer, but nothing more.

61. Ohm, *supra* note 21, at 1751–55.

62. *See infra* Part I.A.

63. *See infra* Part I.B.

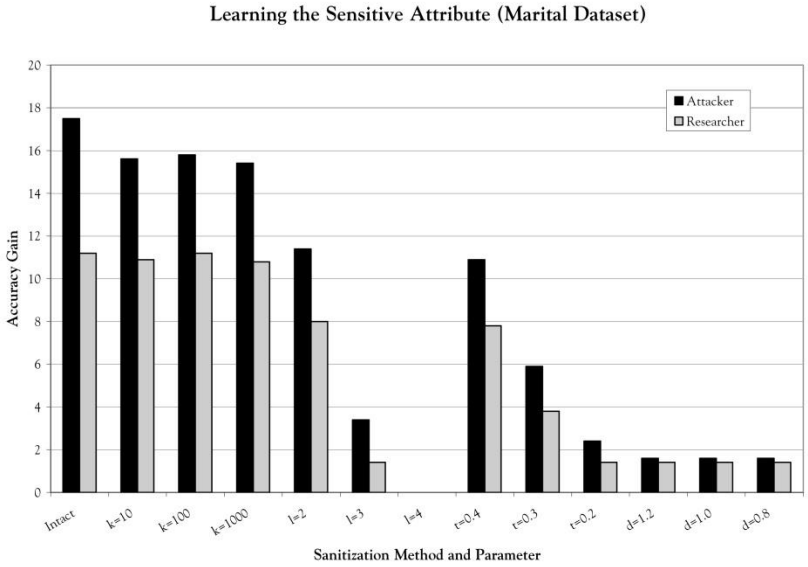
64. Justin Brickell & Vitaly Shmatikov, *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*, 14 PROC. ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 70, 70 (2008).

65. Yakowitz also criticizes Ohm's reliance on the Brickell-Shmatikov paper, similarly pointing out that it is problematic to define privacy and utility to be inverses of one another. Yakowitz, *supra* note 25, at 28–30. She is not correct, however, in asserting that Brickell and Shmatikov "use a definition of data-mining utility that encompasses *all possible* research questions that could be probed by the original database." *Id.* at 30. Brickell and Shmatikov explicitly note that "utility of sanitized databases must be measured empirically, in terms of specific workloads such as classification algorithms," and as described below, their experiments assumed that the researcher had particular classification problems in mind. Brickell & Shmatikov, *supra* note 64, at 74. Nor, as explained below, do I agree with Yakowitz that "the definition of privacy breach used by Brickell and Shmatikov" necessarily "is a measure of the data's utility." Yakowitz, *supra* note 25, at 29.

66. Ohm, *supra* note 21, at 1754.

bars meaning more utility."[67] What is noticeable is the absence of "a short, black bar next to a long, gray bar."[68]

Figure 1



In fact, with a bit more information about what this graph represents, it turns out that it is unsurprising both that the black bars are longer than the gray bars in each pair, and that the two bars largely shrink in proportion to one another across the graph. To understand why requires some additional background on what Brickell and Shmatikov did. Their goal was to measure experimentally the effect of various anonymization techniques on the privacy and utility of data.[69] To do so, they needed to quantify "privacy" and "utility" and then to measure those quantities with respect to a particular research task on a particular data set.[70]

The data set they used was the Adult Data Set from the University of California, Irvine Machine Learning Repository.[71]

67. *Id.*

68. *Id.*

69. Brickell & Shmatikov, *supra* note 64, at 70 ("[W]e measure the tradeoff between privacy (how much can the adversary learn from the sanitized records?) and utility, measured as accuracy of data-mining algorithms executed on the same sanitized records.").

70. *Id.*

71. *See Adult Data Set*, UCI MACHINE LEARNING REPOSITORY, http://archive.ics.uci.edu/ml/datasets/Adult (last visited Feb. 16, 2013).

This is a standard data set that computer scientists have often used to test machine learning theories and algorithms.[72] Extracted from a census database, the data set consists of records that each contain, among other attributes, the age, education, marital status, occupation, race, and sex of an individual.[73]

As is standard in the field, Brickell and Shmatikov defined privacy in an adversarial model, in which privacy is the ability to prevent an "adversary" from learning particular sensitive information.[74] In their model, the adversary is assumed to have some background knowledge about the target individuals, generally in the form of demographic information, such as birth date, zip code, and sex.[75] The goal of anonymization is to prevent the adversary from using the information it already knows to derive sensitive information from the data to be released.[76] For example, a data administrator might want to release medical records in a form that prevents an adversary who knows an individual's birth date, zip code, and sex from finding out about that individual's health conditions.[77] In the experiments that formed the basis for the graph above, the adversary was assumed to know age, occupation, and education, and to be trying to find out marital status.[78]

Brickell and Shmatikov measured a privacy breach by the ability of an adversary to use the background information it already had to determine, or even guess at, the sensitive

72. *See id.* (listing more than fifty papers that cited the data set); *see also* Brickell & Shmatikov, *supra* note 64, at 75 (noting that the authors chose this data set because it had been previously used in other anonymization studies).

73. *See Adult Data Set*, *supra* note 71.

74. *See* Brickell & Shmatikov, *supra* note 64, at 71 ("Privacy loss is the increase in the adversary's ability to learn sensitive attributes corresponding to a given identity.").

75. *See id.* (defining the set Q of quasi-identifiers to be "the set of non-sensitive (e.g., demographic) attributes whose values may be known to the adversary for a given individual").

76. *See id.* at 71–72.

77. *Cf.* Sweeney, *supra* note 22, at 558–59.

78. Brickell and Shmatikov explained that marital status was chosen as the "sensitive" attribute not because of its actual sensitivity in the real world, but because, given the nature of this particular data set, this choice was the best way to maximize the gap between the utility of the data with and without the identifiers known to the adversary. *See* Brickell & Shmatikov, *supra* note 64, at 75 ("We will look at classification of both sensitive and neutral attributes. It is important to choose a workload (target) attribute v for which the presence of the quasi-identifier attributes Q in the sanitized table actually matters. If v can be learned equally well with or without Q, then the data publisher can simply suppress all quasi-identifiers.").

information.[79] Of course, guesses will be right some of the time, even if no data, or only limited data, is released.[80] The measure of privacy loss here was how much *better* the adversary could guess at the sensitive information using the released data than if the data administrator released only the sensitive information, without associating it with any of the information already known to the adversary.[81] In the example above, this means that the baseline for comparison was releasing the data set with the age, occupation, and education fields removed—these were the fields that the adversary was assumed to know. Thus, the "0" line on the graph above, with respect to the black bars, represents the accuracy of the adversary's guesses in this baseline condition, that is, when the data administrator fully suppressed the fields known to the adversary.[82] In this example, that accuracy was 47 percent.[83]

Each of the black bars in the graph above thus represents the privacy loss that resulted from releasing some or all of the

79. *See id.* at 71–72.

80. Even without any released data, an adversary could guess randomly and be right at least some of the time. The fewer choices there are for the sensitive attribute, the more likely a random guess will be correct. For example, if the adversary were trying to determine whether someone does or does not have a particular disease, it could guess randomly and be right at least half the time, because there are only two possible choices. In fact, if the data administrator releases *only* the sensitive information and nothing else, the adversary could at least use that information to determine the frequency of each of the possible choices in the population. For any particular target individual, it could then "guess" that that person has whatever characteristic is most common, and it would be right in proportion to the frequency of that characteristic. So if only 15 percent of the data subjects have a particular disease, then guessing that any one data subject does *not* have the disease is right 85 percent of the time.

81. *See id.* at 76 ("Figure 1 shows the loss of privacy, measured as the gain in the accuracy of adversarial classification A_{acc} . . . ."); *id.* at 73–74 (defining A_{acc} and noting that it "measures the increase in the adversary's accuracy after he observes the sanitized database T' compared to his baseline accuracy from observing T*"); *id.* at 73 (defining T* to be the database in which "all quasi-identifiers have been trivially suppressed").

82. *See id.* at 72 ("The adversary's baseline knowledge A_{base} is the minimum information about sensitive attributes that he can learn after any sanitization, including trivial sanitization which releases quasi-identifiers and sensitive attributes separately.").

83. *Id.* at 76, fig. 1 ("With trivial sanitization, accuracy is 46.56 [percent] for the adversary . . . ."). There were seven possible values for the sensitive attribute, marital status. *See Adult Data Set*, *supra* note 71. Guessing randomly would thus produce an accuracy of 1/7, or approximately 14 percent. Apparently, however, 47 percent of the population shared the most common marital status. An adversary who sees only the marital status column of the database could therefore guess correctly as to any one individual 47 percent of the time.

information about age, occupation, and education.[84] The leftmost bar corresponds to the full disclosure of the data set.[85] At a value of about 17, this means that an adversary who knew the age, occupation, and education of a target individual, and was given the complete data set, would have been able to guess that person's marital status correctly 64 percent of the time.[86]

The remaining bars correspond to the release of "anonymized" data.[87] In particular, Brickell and Shmatikov subjected the data to the techniques of generalization and suppression.[88] Suppression means entirely deleting certain fields in the database.[89] In generalization, a more general category replaces more specific information about an individual.[90] "City and state" could be generalized to just "state" alone. Race could be generalized to "white" and "non-white." Age could be generalized to five-year bands. In this way, an adversary looking for information about a 36-year-old Asian person whose zip code is 10003, for example, would know only that the target record is among the many records of non-whites between the ages of 36 and 40 from New York state. The shrinking black bars represent the fact that as more of the age, occupation, and education information was generalized, the adversary's ability to guess marital status shrank back toward the baseline level.

As for defining utility, Brickell and Shmatikov specified a particular task that a hypothetical researcher wanted to perform on the data set.[91] Utility could then be measured by how well the researcher could perform the task, given either the full data set or some anonymized version of it.[92] In this paper, Brickell and Shmatikov were interested in the usefulness of anonymized data for data mining, and, in particular, for the task of building "classifiers."[93] A classifier is

84. *See id.* at 76.

85. *See id.*

86. *See id.* This is the 47 percent baseline accuracy plus the 17 percent height of the leftmost bar.

87. *See id.* at 72–73 (noting that the forms of privacy tested were k-anonymity, l-diversity, t-closeness, and δ-disclosure privacy, and defining each of these).

88. *See id.* at 72.

89. *See* Ohm, *supra* note 21, at 1713–14.

90. *See id.* at 1714–15.

91. Brickell & Shmatikov, *supra* note 64, at 75 ("We must also choose a workload for the legitimate researcher.").

92. *See id.*

93. *See id.* at 74.

a computer program that tries to predict one attribute based on the value of other attributes.[94] For example, a researcher might want to build a program that could predict whether someone will like the movie *The Lorax* based on this person's opinion of other movies. The idea is to use a large data set in order to build such a classifier in an automated way by mining the data for patterns, rather than using human intuition to hypothesize, for example, that those who enjoyed *Horton Hears a Who* might also enjoy *The Lorax*.

A classifier built using anonymized data will generally be less accurate than one built using the original data.[95] Generalization hides patterns that become contained entirely within a more general category. If residents of Buffalo and New York City have very different characteristics—suppose one group likes *The Lorax* much more than the other group—this will be obscured if both groups are categorized as residents of New York state. So, for example, a classifier that has access to full city information will tend to be more accurate than one that only knows state information.

The gray bars in the graph above show the utility of the different data sets, that is, the accuracy of a classifier built using each data set.[96] Again, the leftmost bar indicates the utility of the full data set, while the other bars indicate the utility of various anonymized data sets.[97] Importantly, Brickell and Shmatikov used the same baseline condition for the privacy bars as for the utility bars, namely, the data set with the age, occupation, and education fields removed.[98] The gray bars thus plot the gain in utility when the researcher has at least some access to age, occupation, and education information, as compared to when she has no access to this information at all.

Recall, however, that the hypothetical researcher was trying to construct a classifier and that the goal of a classifier is to predict one of the attributes, given the other attributes. Which attribute was the researcher's classifier trying to predict

94. *See generally* TOM MITCHELL, MACHINE LEARNING (1997).

95. *See* Brickell & Shmatikov, *supra* note 64, at 75.

96. *See id.* at 76.

97. *See id.*

98. *See id.* (explaining that the graph compares the privacy loss to "the gain in workload utility $U_{san} - U_{base}$"); *id.* at 74 (explaining that U_{base} is computed by picking "the trivial sanitization with the largest utility" and that "trivially sanitized datasets" are "datasets from which either all quasi-identifiers Q, or all sensitive attributes S have been removed").

in the experiments graphed above? In fact, it was marital status,[99] *precisely the sensitive attribute that the data administrator was simultaneously trying to hide from the adversary*. In this particular experiment, privacy loss was measured by the adversary's ability to guess marital status, and utility was measured by the researcher's ability to guess marital status using the very same data. It should come as no surprise then that *so defined*, it was impossible to achieve privacy and utility at the same time. Any given anonymization technique either made it more difficult to predict marital status or it did not. The black and gray bars thus naturally maintained roughly the same proportion to each other, no matter what technique was used.[100]

Brickell and Shmatikov actually recognized this limitation in the experiments graphed above, noting that "[p]erhaps it is not surprising that sanitization makes it difficult to build an accurate classifier for the sensitive attribute."[101] They went on to describe the results of experiments in which the researcher and adversary were interested in different attributes.[102] These results are somewhat ambiguous. The graph reproduced below as Figure 2, for example, appears to show several examples in which the leftmost bar in the set has shrunk significantly (i.e., the released data set is significantly more private), while the remaining bars have not shrunk much (i.e., not much utility has been lost).[103] Brickell and Shmatikov do not discuss the

99. *See id.* at 76, fig. 1 ("Gain in classification accuracy for the sensitive attribute (marital) in the 'Marital' dataset.").

100. Nor is there any significance to the fact that the black bars are always longer than the gray bars. Both the adversary's gain and the researcher's gain were measured relative to the baseline condition in which the *adversary's* additional information had been suppressed. *See id.* at 74. In that baseline condition, the adversary would be guessing randomly, while the researcher would have access to the remaining information in the data set and could thus do better. In the example graphed above, the researcher's accuracy in the baseline condition was 58 percent, compared to 47 percent for the adversary. *Id.* at 76 fig. 1. This means that the "0" line in the graph represents an accuracy of 58 percent with respect to the gray bars, but 47 percent with respect to the black bars. Relative to their respective baselines, one would expect the adversary to have more to gain from having at least some age, occupation, and education information than the researcher, because the adversary is going from nothing to something, whereas the researcher is only adding to the information she already had. Naturally then, the black bars are longer than the gray bars.

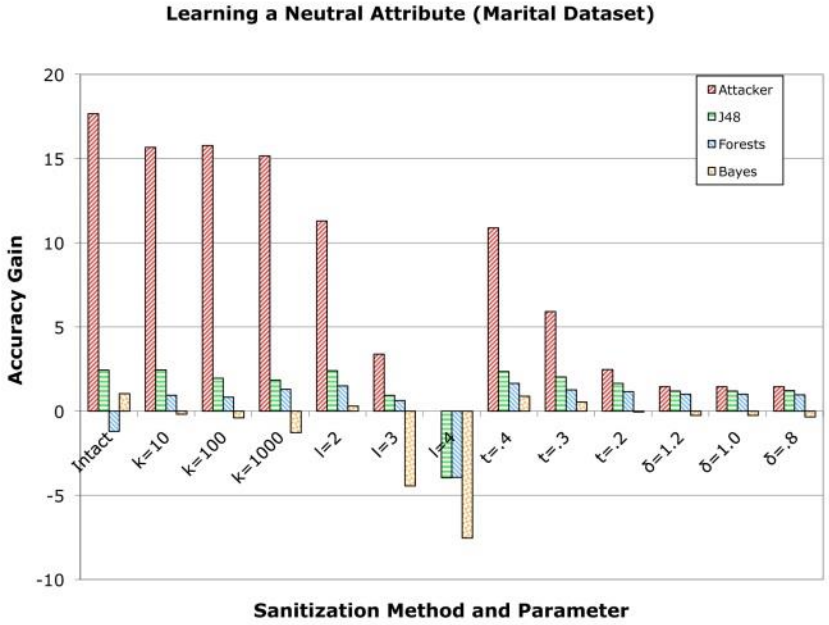
101. *Id.*

102. *Id.* ("We now consider the case when the researcher wishes to build a classifier for a *non-sensitive* attribute *v*.").

103. *Id.* at 77, fig. 3 ("Gain in the adversary's ability to learn the sensitive attribute (marital) and the researcher's ability to learn the workload attribute

implications of this particular graph in their paper.[104]

Figure 2



The lesson here is that the meaning of a broad claim like "even modest privacy gains require almost complete destruction of the data-mining utility"[105] can only be understood with respect to particular definitions of "privacy" and "utility." In the example that Ohm uses, privacy and utility were essentially defined to be inverses of one another, because the privacy goal was aimed at hiding exactly the information that the researcher was seeking.[106] So defined, we should not be surprised to find that we cannot achieve both privacy and utility simultaneously, but such a result does not apply to other reasonable definitions of privacy and utility.[107]

(salary) for the 'Marital' dataset."). In this experiment, the authors tested three "different machine learning algorithms" for constructing classifiers. *Id.* at 76. Hence, there are three "utility" bars in each set. Again, what matters is the length of the bars relative to the corresponding one in the first set, not their lengths relative to the others in the same set. *See supra* note 100 and accompanying text.

104. *See id.* at 75.

105. *Id.* at 70.

106. *See supra* note 99 and accompanying text.

107. *See, e.g.*, Noman Mohammed et al., *Differentially Private Data Release for*

To be sure, the experiments documented in the first graph above confirm something important about the relationship between privacy and utility: what is good for the goose (the data-mining researcher) is good for the gander (the adversary). Thus, when the point of the research is to study a sensitive characteristic, we will need to consider carefully whether to regard any of the data available to the researcher as potentially privacy-invading.[108] Such a study does not, however, establish that privacy and utility will inevitably conflict in all contexts.

B. Differential Privacy

Techniques to achieve a concept called "differential privacy" might also be more helpful than Ohm's article suggests. The motivation for the concept of differential privacy is captured by the following observation: in the worst case, it is always theoretically possible that any information revealed by a data set is the missing link that the adversary needs to breach someone's privacy.[109] For example, if the adversary is trying to learn someone's height and knows that it is exactly two inches shorter than the height of the average Lithuanian woman, then a data set that reveals the height of the average Lithuanian woman allows the adversary to learn the target information.[110]

In such a situation, however, one might naturally attribute the privacy breach to the adversary's prior knowledge, rather than to the information revealed by the data set. Intuitively, while the revelation of the data set was a cause-in-fact of the privacy breach, it was not a proximate cause. To make sense of this intuition, notice that the information revealed by the data set, about the average height of a Lithuanian woman, would be approximately the same whether or not the target individual

---

Data Mining, 17 PROC. ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 493, 494 (2011) ("We present the first generalization-based algorithm for differentially private data release that preserves information for classification analysis."). For a definition of "differential privacy," see *infra* Part I.B.

108. *See infra* Part III.C.

109. Cynthia Dwork, who originated the concept of differential privacy, formalizes this intuition and gives a proof. *See* Cynthia Dwork, *Differential Privacy*, 33 PROC. INT'L COLLOQUIUM ON AUTOMATA LANGUAGES & PROGRAMMING 1, 4–8 (2006).

110. This is the example that Dwork gives. *See id.* at 2; *see also* Ohm, *supra* note 21, at 1752.

appeared in the data set. In order for the computed average to be accurate, it must have been based on the information of many people, so that the target person's presence or absence in the data set would not significantly affect the overall average, even if the target person were herself a Lithuanian woman. The goal of differential privacy is thus to reveal only information that does not significantly depend on any one individual in the data set.[111] In this way, any negative effects that an individual suffers as a result of the data release are ones that cannot be traced to the presence of her data in the data set.[112]

Dwork shows that it is possible to achieve differential privacy in the "interactive" setting, in which the data administrator answers questions about the data, but never releases even a redacted form of the entire data set.[113] Rather than answer the researcher's questions exactly, the data administrator adds some random noise to the answers, changing them somewhat from the true answers.[114] The amount of noise depends on the extent to which the answer to the question changes when any one individual's data changes.[115] 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.[116]

Contrary to Ohm's characterization, however,[117] differential privacy has also been studied in the "non-interactive" setting, in which some form of data *is* released, without any need for further participation by the data

111. *See* Dwork, *supra* note 109, at 2.

112. *See* Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMM. OF THE ACM 86, 91 (2011).

113. *See* Dwork, *supra* note 109, at 9–11.

114. *See id.* at 9–10; *see also* Cynthia Dwork et al., *Calibrating Noise to Sensitivity in Private Data Analysis*, 3 PROC. THEORY OF CRYPTOGRAPHY CONF. 265 (2006).

115. *See* Dwork, *supra* note 109, at 10.

116. *See id.*

117. *See* Ohm, *supra* note 21, at 1755–56 (describing differential privacy as an "interactive technique" and noting that interactive techniques "tend to be less flexible than traditional anonymization" because they "require constant participation from the data administrator").

administrator.[118] It is true that more questions can be answered in a differentially private way with an interactive mechanism than with a non-interactive data release.[119] At least some non-trivial questions can be answered in the non-interactive setting, however, and computer scientists may yet discover ways to do more.[120] Thus, whether these techniques are too limited can only be evaluated with respect to the particular uses that a researcher might have in mind, or in other words, only with respect to a particular conception of utility. At least in some domains, with some research questions, non-interactive techniques can provide both differential privacy and a form of utility.

Ohm also incorrectly suggests that differential privacy techniques are "of limited usefulness" simply because they require the addition of noise.[121] The noise added by a differential privacy mechanism, however, is calibrated by design to drown out information about specific individuals, while affecting more aggregate information substantially less.[122] Ohm cites an example in which police erroneously and repeatedly raided a house on the basis of noisy data,[123] but this example shows only that the noise-adding techniques were doing their job. Noise is supposed to make the data unreliable with respect to any one individual, and, thus, the problem in that example is not that noise was added to the data, but that police were using noisy data to determine which search

118. *See generally* Avrim Blum et al., *A Learning Theory Approach to Non-Interactive Database Privacy*, 40 PROC. ACM SYMP. ON THEORY OF COMPUTING 609 (2008); Cynthia Dwork et al., *On the Complexity of Differentially Private Data Release*, 41 PROC. ACM SYMP. ON THEORY OF COMPUTING 381, 381 (2009) ("We consider private data analysis in the setting in which a trusted and trustworthy curator, having obtained a large data set containing private information, releases to the public a 'sanitization' of the data set that simultaneously protects the privacy of the individual contributors of data and offers utility to the data analyst."); Cynthia Dwork et al., *Boosting and Differential Privacy*, 51 PROC. IEEE SYMP. ON FOUND. OF COMPUTER SCI. 51 (2010); Moritz Hardt et al., *Private Data Release via Learning Thresholds*, 23 PROC. ACM-SIAM SYMP. ON DISCRETE ALGORITHMS 168 (2012).

119. *See* Blum et al., *supra* note 118, at 616–17.

120. *See id.* at 615 (stating as a significant open question the extent to which it is "possible to *efficiently*[,] privately[,] and usefully release a database" that can answer a wider variety of questions).

121. *See* Ohm, *supra* note 21, at 1757.

122. *See supra* notes 114–116 and accompanying text.

123. *See* Ohm, *supra* note 21, at 1757 (citing *Cops: Computer Glitch Led to Wrong Address*, MSNBC NEWS (Mar. 19, 2010), http://www.msnbc.msn.com/id/35950730/ns/us_news-crime_and_courts/t/cops-computer-glitch-led-wrong-address/ (last visited Jan. 26, 2013)).

warrants to obtain and which houses to raid. Those tasks involve singling out individuals and are not the sort of aggregate purposes to which differential privacy or other noise-adding techniques are suited. Certainly some socially useful research might require non-noisy data,[124] but the use of noise should not be regarded as an inherent problem in all contexts.

In literature that Ohm does not cite, computer scientists have indeed proved some fundamental limits on the ability to release data while still protecting privacy.[125] In particular, getting answers to too many questions about arbitrary sets of individuals in a sensitive data set allows an adversary to reconstruct virtually the entire data set, even if the answers he or she gets are quite noisy.[126] However, a system that either answers fewer questions or only answers questions of a particular form can be differentially private.[127] Thus, as Part I.A also demonstrated with respect to the Brickell-Shmatikov paper, the proven limits in the computer science literature are only limits with respect to particular definitions of privacy and utility, definitions that may apply in some contexts, but not all.

II. WHY WE SHOULDN'T BE TOO OPTIMISTIC ABOUT ANONYMIZATION

Jane Yakowitz criticizes Ohm and others for overstating the risk of re-identification and under-appreciating the value of public data releases.[128] She proposes that the law ought to be

124. *See infra* notes 280–285 and accompanying text.

125. *See, e.g.*, Irit Dinur & Kobbi Nissim, *Revealing Information While Preserving Privacy*, 22 PROC. ACM SYMP. ON PRINCIPLES OF DATABASE SYSTEMS 202 (2003).

126. *See id.* at 204 ("[W]e show that whenever the perturbation is smaller than $\sqrt{n}$, a polynomial number of queries can be used to efficiently reconstruct a 'good' approximation of the entire database."); *see also id.* ("We focus on binary databases, where the content is of n binary (0-1) entries . . . . A statistical query specifies a subset of entries; the answer to the statistical query is the number of entries having value 1 among those specified in it.").

127. *See* Blum et al., *supra* note 118, at 610 ("We circumvent the existing lower bounds by only guaranteeing usefulness for queries in restricted classes.").

128. *See* Yakowitz, *supra* note 25, at 4. Yakowitz's paper is often cited on the opposite side of the debate from Ohm's. *See, e.g.*, Pamela Jones Harbour et al., Sorrell v. IMS Health Inc.*: The Decision and What It Says About Patient Privacy*, FULBRIGHT & JAWORSKI L.L.P. (June 30, 2011), http://www.fulbright.com/index.cfm?FUSEACTION=publications.detail&PUB_ID=5000&pf=y ("Professor Ohm has warned that increases in the amount of data and advances in the technology used to analyze it mean that data can be de-anonymized . . . . Others, however, such as Jane Yakowitz, . . . have downplayed the risk of such de-anonymization.").

encouraging data release, not discouraging it, and she argues that there should be a safe harbor for the disclosure of data that has been anonymized using relatively straightforward techniques.[129] While Ohm's conceptions of privacy and utility may be too broad to apply to all contexts, Yakowitz's conceptions may be too narrow. In particular, Yakowitz's reliance on the concept of "*k*-anonymity," as well as her citation to particular studies of re-identification risk, are both premised on a particular conception of what counts as a privacy violation and what counts as a useful research result.

A. k-*Anonymity*

Yakowitz essentially argues that the concept of "*k*-anonymity" sufficiently captures the privacy interest in data sets, and that imposing *k*-anonymity as a requirement for data release will largely preserve the utility of data sets, while posing only a minimal privacy risk.[130] The concept of *k*-anonymity originated with the work of Latanya Sweeney, who demonstrated, rather vividly, that birth date, zip code, and sex are enough to uniquely identify much of the U.S. population.[131]

---

129. Yakowitz, *supra* note 25, at 44–47.

130. Yakowitz calls this ensuring a "minimum subgroup count." *Id.* at 45. She also states that, in the alternative, the data producer can ensure an "unknown sampling frame," which means that an adversary cannot tell whether any given individual is in the data set or not. *Id.* In fact, the two possible requirements are computationally equivalent. If the adversary does not know whether the target individual is in the data set or not, then one can imagine replacing the actual sampled data set with the complete data set from which it was drawn (and in which the adversary is sure the subject is present). The complete data set can be thought of as having an extra field that indicates whether the subject was in the original sampled data set. In this situation, this is simply another field as to which the adversary happens to lack information. The adversary's ability to isolate a set of matching records in this master data set then corresponds to its ability to learn something about the target individual in the original data set. In this sense, an unknown sampling frame, while making it easier to satisfy *k*-anonymity because the relevant data set has effectively been expanded, does not obviate the need to guarantee *k*-anonymity at some level. Yakowitz implicitly acknowledges this in conceding that the requirement of unknown sampling frame can fail to protect privacy "in circumstances where a potential data subject is unusual." *See* Yakowitz, *supra* note 25, at 46 n.230.

131. *See* Sweeney, *supra* note 22, at 558; *see also* Philippe Golle, *Revisiting the Uniqueness of Simple Demographics in the US Population*, 5 PROC. ACM WORKSHOP ON PRIVACY IN THE ELEC. SOC'Y 77, 77 (2006) (revisiting Sweeney's work and finding that birth date, zip code, and sex uniquely identified 61 percent of the U.S. population in 1990, as compared to Sweeney's finding of 87 percent). Sweeney used this information to pick out then-Governor Weld's medical records from a database released by the state of Massachusetts. *See* Sweeney, *supra* note

Thus, an adversary who knows these three pieces of information about a target individual can likely pick out that person's record from a database that contains these identifiers. More generally, given the identifiers known to the adversary, we can imagine the adversary searching the database for all matching records. For example, if the adversary knows that the target person is a white male living in zip code 10003, and race, sex, and zip code fields appear in the database, then the adversary can collect the records that match those fields and determine that the target individual's record is one of them.[132] If there is only one matching record, then the adversary will have identified the target record exactly. The concept of *k*-anonymity requires the data administrator to ensure that, given what the adversary already knows, the adversary can never narrow down the set of potential target records to fewer than *k* records in the released data.[133] This guarantee is generally accomplished through suppression and generalization, as described above.[134]

The trouble with relying on *k*-anonymity as the sole form of privacy protection is that it has some known limitations. The first is that it may be possible to derive sensitive information from a database without knowing precisely which record corresponds to the target individual.[135] For example, if the adversary is able to narrow down to a set of records that all share the same sensitive characteristic, then he will have determined that the target individual has this sensitive characteristic. Suppose there are ten white males on one particular city block, and one of them is the target individual. If a database shows that all ten of these men have hypertension, then the adversary would be able to learn something about the target individual from the database, even without being able to determine which of the ten records is the target. More generally, if eight out of these ten men have

22, at 559.

132. This discussion assumes that the adversary knows whether or not a given person is in the database. If not, see *supra* note 130.

133. *See* Sweeney, *supra* note 22, at 564–65.

134. *See* Latanya Sweeney, *Achieving* k-*Anonymity Privacy Protection Using Generalization and Suppression*, 10 INT'L J. UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYSTEMS 571 (2002); *see also* Ohm, *supra* note 21, at 1713–15; *supra* notes 88–90 and accompanying text.

135. This is known in the literature as a "homogeneity attack." *See* Ashwin Machanavajjhala et al., *L-Diversity: Privacy Beyond* k-*Anonymity*, 1 ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA, Article 3, 3–4 (2007).

hypertension, for example, the adversary would be able to make a much better guess about the hypertension status of the target person than he was able to make before the data was released.

Yakowitz's answer to this problem is that if a particular demographic group indeed shares a particular sensitive characteristic, then this is a research result that ought to be publicly available, not a private fact to be hidden.[136] Whether such information should be regarded as legitimate research, however, depends heavily on context. Certainly the fact that women in Marin County, California had a high incidence of breast cancer is of significant public health interest,[137] even though the disclosure of this fact improves others' ability to guess whether any particular woman living in Marin County had breast cancer. Suppose instead that a database discloses that one out of the ten men over forty on a particular suburban block is HIV-positive. Such a fact would seem to have no research significance,[138] while potentially exposing the men on that block to privacy harms.[139]

Another limitation of *k*-anonymity is that its privacy guarantees depend heavily on knowing what background information the adversary already has.[140] If the adversary turns out to have more than expected, then he may be able to leverage this information to discover additional sensitive information from the released data. For example, the released data might ensure that basic demographic information could be used only to narrow the set of potential medical records down to a set of five or more. Perhaps an adversary who knows the month and year of a hospital admission, however, would be able to pick out the target record from among those with the same demographic characteristics.

---

136. *See* Yakowitz, *supra* note 25, at 29. Yakowitz does acknowledge the potential problem and suggests that "[a]dditional measures may be taken if a subgroup is too homogenous with respect to a sensitive attribute." *Id.* at 54 n.262. She does not, however, appear to require any such measures in the safe harbor she proposes, nor does she consider the implications of such a requirement on the utility of the resulting data. *See id.* at 44–46.

137. *See* Christina A. Clarke et al., *Breast Cancer Incidence and Mortality Trends in an Affluent Population: Marin County, California, USA, 1990–1999*, 4 BREAST CANCER RESEARCH R13 (2002), *available at* http://breast-cancer-research.com/content/4/6/R13.

138. *See infra* Part III.C.

139. *See infra* Part III.B.

140. This is a "background knowledge attack." *See* Machanavajjhala et al., *supra* note 135, at 4–5.

Yakowitz draws the line at information that is "systematically compiled and distributed by third parties,"[141] and would impose no requirement to hide sensitive information from an adversary who has additional background knowledge. Such a view assumes that privacy protections in this setting are primarily directed against strangers, people who have no inside information that they can leverage. As further developed below, the view that privacy law is intended to protect only against outsiders and not insiders is one that may be appropriate for some contexts, but not for others.[142]

B. Re-identification Studies

Yakowitz also relies on studies that suggest that the "realistic" rate of re-identification is quite low.[143] A recent example of a paper in this vein is the meta-study conducted by Khaled El Emam and others.[144] They surveyed the literature to find reported re-identification attacks, and while overall they found that studies reported a relatively high re-identification rate, they downplayed the significance of many of these studies, finding only two "where the original data was de-identified using current standards."[145] Of those two studies, only one "was on health data, and the percentage of records re-identified was 0.013 [percent], which would be considered a very low success rate."[146]

Whether such studies are in fact an appropriate measure of privacy risk, however, again depends on how one conceives of privacy. Both the El Emam meta-study and the Lafky study measured the risk that individual records could be re-identified, that is, associated with the name of the individual whose record it was.[147] Indeed, the El Emam study looked to

141. Yakowitz, *supra* note 25, at 45.

142. *See infra* Part III.A.

143. *See* Yakowitz, *supra* note 25, at 28 (citing DEBORAH LAFKY, DEP'T OF HEALTH & HUMAN SERVS., THE SAFE HARBOR METHOD OF DE-IDENTIFICATION: AN EMPIRICAL TEST (2009)); *see also* Peter K. Kwok & Deborah Lafky, *Harder Than You Think: A Case Study of Re-identification Risk of HIPAA-Compliant Records* (2011).

144. *See* Khaled El Emam et al., *A Systematic Review of Re-Identification Attacks on Health Data*, 6 PLoS ONE e28071 (2011), *available at* http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0028071.

145. *Id.* at 8–9.

146. *Id.* at 9. The referenced study was again that of Kwok and Lafky. *See id.* at 6 (citing Kwok & Lafky, *supra* note 143).

147. *See id.* at 2; Kwok & Lafky, *supra* note 143, at 2 ("[O]ur model of intrusion

see whether re-identifications were verified, stating that a re-identification attack should not be regarded as successful "unless some means have been used to verify the correctness of that re-identification."[148] The authors regarded verification as necessary "[e]ven if the probability of a correct re-identification is high."[149]

As previously described, not every arguable privacy breach requires the adversary to match records to identities. An adversary may be able to learn sensitive information about a particular individual even if the adversary cannot determine which record belongs to that individual.[150] The El Emam study did not include such potential attacks in its model of a privacy violation.[151]

Moreover, the El Emam and Lafky studies did not consider whether what they regarded as appropriate de-identification might significantly degrade the utility of the data set. Both studies looked for re-identification attacks against data sets that had been de-identified using "existing standards,"[152] in particular, the Safe Harbor standard specified in the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.[153] That standard specifies a list of eighteen data elements that must be suppressed or generalized, including the last two digits of zip codes and all dates except years.[154] Such a standard potentially goes well beyond the *k*-anonymity rule advocated by Yakowitz.[155]

Suppression of zip code digits, exact dates, and other such data, however, can make the data significantly less useful for certain tasks. Almost all of Manhattan shares the same first three zip code digits.[156] Thus, any study designed to look for differences within Manhattan could not be conducted using

---

focused on only identity disclosure.").

148. El Emam et al., *supra* note 144, at 3.

149. *Id.*

150. *See supra* note 135 and accompanying text.

151. *See* El Emam et al., *supra* note 144, at 2.

152. *Id.* at 3.

153. *See* Kwok & Lafky, *supra* note 143, at 2.

154. *See* 45 C.F.R. § 164.514(b)(2) (2013).

155. For example, there may be thousands of people in each zip code, such that a database that keyed information only to zip code might be *k*-anonymous for some large *k*. Nevertheless, the HIPAA Safe Harbor standard would require the suppression of at least the last two digits of the zip codes.

156. *See ZIP Code Definitions of New York City Neighborhoods*, N.Y. STATE DEP'T OF HEALTH (Mar. 2006), http://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm.

safe harbor data. Similarly, studies looking for trends within the same year would not be possible. For example, tracing trends relative to the 2012 presidential election campaign would be impossible, because all of the events of interest occurred within a single year.[157]

Implicit in these re-identification studies then is a conception of utility that excludes certain types of research. Moreover, both these studies and the *k*-anonymity model implicitly adopt a view of privacy that does not protect against certain intrusions, such as an adversary discovering an individual's sensitive information without identifying that individual's record in the data set. These implicit choices about how to define privacy and utility may be appropriate in some contexts, but one should not assume that they apply across all contexts.

III. THE CONCEPTS OF PRIVACY AND UTILITY

As Parts I and II have shown, advocates and detractors of anonymization have very different conceptions about what "privacy" and "utility" mean, and consequently, they have come to very different conclusions about the relationship between privacy and utility. To begin to bridge the gap between the opposing sides of this debate, and to guide policymakers, what is needed is a clearer understanding of how and why conceptions of privacy and utility vary. Accordingly, this Part develops a framework for analyzing conceptions of privacy and utility. With such a framework, policymakers will better understand what is at stake in competing calls for greater or lesser privacy protection in data sets, and they will be better able to craft solutions appropriate to the specific contexts in which the problem arises.

With respect to defining privacy, a key insight is that varying conceptions of privacy can be traced to varying conceptions of the threats against which individuals need protection. Part III.A explores the concept of "privacy threats" and the need to specify what information should be hidden and from whom, before we can address what legal or technical tools to use to accomplish these goals. Moreover, as described in Part III.B, data release often results in the disclosure of information

157. It was the fact that dates were included in the data set that made the Netflix Prize data set fail the safe harbor standard. *See* El Emam et al., *supra* note 144, at 7.

about individuals that is not known with certainty. Whether to treat the disclosure of such uncertain information as a privacy breach also depends heavily on what harms we ultimately seek to prevent.

On the utility side of the equation, Part III.C demonstrates that the legitimacy of what might be called research is highly contextual and a potential source of disagreement. These disagreements matter for whether de-identification is an effective privacy tool because, as we will see, some types of research are harder to accomplish privately than others. Finally, Part III.D points out that utility has an important temporal dimension, and the extent to which we want to support future unpredictable uses of data will greatly influence the level of privacy that we can obtain.

A. Privacy Threats

The idea that the term "privacy" is heavily overloaded is by now well established.[158] It can be used to name a wide variety of concepts, norms, laws, or rights, ranging from the "right to be let alone"[159] to a respect for "contextual integrity."[160] In the context of data release, it might seem at first glance that this definitional problem can be avoided. All perhaps agree that the relevant privacy goal here is that of hiding one's identity. As we have seen, though, different scholars have very different ideas about what it means to hide one's identity.

The computer science literature provides a model for how the law can and should make these differences explicit. To a computer scientist, privacy is defined not by what it is, but by what it is not—it is the absence of a privacy breach that defines a state of privacy.[161] Defining privacy thus requires defining what counts as a privacy breach, and to do that, the computer scientist imagines a contest between a mythical "adversary" and the designer of the supposedly privacy-preserving system.[162] The adversary has certain resources at his disposal,

158. *See generally* SOLOVE, *supra* note 20.

159. *See* Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 193 (1890).

160. *See generally* HELEN FAY NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010).

161. *See, e.g.*, Dwork, *supra* note 109, at 1 (defining privacy by asking "What constitutes a failure to preserve privacy?").

162. *Id.* ("What is the power of the adversary whose goal it is to compromise privacy?").

including prior knowledge, computational power, and access to the data set. The adversary is then imagined as trying to attack the private system and accomplish some specified goal. If the adversary can succeed at its goal, then we say that the system fails to protect privacy. If the adversary fails, then the system succeeds.

To give content to the concept of privacy that we are seeking to protect, we must therefore specify the nature of the adversary we are protecting against. This includes specifying the adversary's goals, specifying the tools available to the adversary and the ways in which it can interact with the protected data, and specifying the adversary's capabilities, both in terms of computational power or sophistication and in terms of the background information that the adversary has before interacting with the protected data. Specifying each of these is necessary to give meaning to a claim that de-identification either succeeds or fails at protecting privacy in a given context.

Stated differently, we need to define the threats that de-identification is supposed to withstand. Long made explicit in the area of computer security,[163] threat modeling is equally important with respect to analyzing data privacy.[164] Different commentators and researchers have had different privacy threats in mind and have, therefore, come to different conclusions about the effectiveness of de-identification. Should we worry about the colleagues we talk to around the "water cooler"?[165] Or should we focus only on "the identity thief and the behavioral marketer"?[166] The question is important because the scope of the threats we address determines the scope of the privacy protection we obtain. Thinking in terms of threats focuses the policy discussion and guides policymakers more directly to address three steps: identifying threats, characterizing them, and then crafting policy solutions to

163. *See* BRUCE SCHNEIER, SECRETS AND LIES: DIGITAL SECURITY IN A NETWORKED WORLD 12–22 (2000); *see also* SUSAN LANDAU, SURVEILLANCE OR SECURITY?: THE RISKS POSED BY NEW WIRETAPPING TECHNOLOGIES 145–73 (2010).

164. For a recent example of the beginnings of privacy threat modeling, see Mina Deng et al., *A Privacy Threat Analysis Framework: Supporting the Elicitation and Fulfillment of Privacy Requirements*, 16 J. REQUIREMENTS ENGINEERING 3, 3 (2011) ("Although digital privacy is an identified priority in our society, few systematic, effective methodologies exist that deal with privacy threats thoroughly. This paper presents a comprehensive framework to model privacy threats in software-based systems.").

165. Narayanan & Shmatikov, *supra* note 8, at 122.

166. Yakowitz, *supra* note 25, at 39.

address those threats.

1. Identifying Threats: Threat Models

The term "threat model" is used in computer security in at least two distinct ways. On the one hand, threat modeling can describe the activity of systematically identifying who might try to attack the system, what they would seek to accomplish, and how they might carry out their attacks.[167] For example, in evaluating the security of a password-protected banking website, one would want to consider the possibility of an intruder stealing money from customer accounts by intercepting information flowing to and from the website, guessing customers' passwords, or perhaps infecting customers' computers with a virus that logged their keystrokes.

On a different view, the "threat model" of a security system is the set of threats that the system is designed to withstand.[168] Ideally, of course, those threats are identified through a process of threat modeling so that the design of the system matches up to the reality of the threats in the world. Systems have threat models in this second sense, however, regardless of whether those models have been made explicit and regardless of whether they fit with reality.[169]

Privacy laws, no less than privacy technologies, have such implicit threat models. That is, any given privacy law addresses certain types of privacy invasions, but not others. And just as with privacy technologies, there can be a mismatch between the implicit threat model in the law and the reality in the world.

For example, consider the case of *United States v. Councilman*.[170] Brad Councilman was vice president of

167. *See* MICHAEL HOWARD & DAVID LEBLANC, WRITING SECURE CODE 69 (2d ed. 2003) ("A threat model is a security-based analysis that helps people determine the highest level security risks posed to the product and how attacks can manifest themselves.").

168. *See, e.g.*, Derek Atkins & Rob Austein, *RFC 3833—Threat Analysis of the Domain Name System (DNS)*, THE INTERNET ENGINEERING TASK FORCE (2004), http://tools.ietf.org/html/rfc3833 (stating as its goal the documentation of "the specific set of threats against which DNSSEC [the Domain Name System Security Extensions] is designed to protect").

169. *See* SCHNEIER, *supra* note 163, at 12 (noting that the design of a secure system involves "conscious or *unconscious* design decisions about what kinds of attacks . . . to prevent . . . and what kinds of attacks . . . to ignore") (emphasis added).

170. 418 F.3d 67 (1st Cir. 2005) (en banc).

Interloc, an online rare book listing service.[171] Interloc worked with book dealers to list and sell those dealers' books to the public. As part of this business relationship, Interloc provided e-mail addresses in the Interloc.com domain to its affiliated book dealers and acted as the service provider for these e-mail services.[172] According to the indictment against him,[173] Councilman directed that e-mails sent to book dealer accounts from Amazon.com be copied and stored for him and other Interloc employees to read, ostensibly to obtain a competitive advantage over Amazon.[174]

Councilman was charged with violating the Wiretap Act.[175] The district court held that acquiring communications in "electronic storage," as these e-mails were, was not an interception of "electronic communications" within the meaning of the Wiretap Act.[176] A panel of the First Circuit initially affirmed,[177] but the court later granted rehearing en banc and reversed, holding that communications in electronic storage are within the scope of the Wiretap Act.[178]

The Electronic Communications Privacy Act, however, has another section that seemingly would have been a better fit from the start for Councilman's actions. The Stored Communications Act (SCA) prohibits unauthorized access to communications in electronic storage.[179] Why didn't the government simply fall back on charging a violation of the SCA?

The trouble is that, while the SCA prohibits a service provider from *disclosing* the contents of communications,[180] it contains an explicit exception for *access* to those

171. *Id.* at 70.

172. *Id.*

173. The court considered the facts as alleged in the indictment because the case had been decided on a motion to dismiss. *Id.* at 71–72. A jury later acquitted Councilman. *See* Stephanie Barry, *Jury Acquits Ex-Selectman of Conspiracy*, THE REPUBLICAN, Feb. 7, 2007, at A1.

174. *Councilman*, 418 F.3d at 70–71.

175. *See* 18 U.S.C. § 2511 (2012).

176. United States v. Councilman, 245 F. Supp. 2d 319 (D. Mass. 2003), *vacated and remanded*, 418 F.3d 67 (1st Cir. 2005) (en banc).

177. United States v. Councilman, 373 F.3d 197 (1st Cir. 2004).

178. *Councilman*, 418 F.3d at 72.

179. 18 U.S.C. § 2701 (2012).

180. 18 U.S.C. § 2702(a). There are various exceptions, including one for disclosures "as may be necessarily incident to the rendition of the service or to the protection of the rights or property of the provider of that service," 18 U.S.C. § 2702(b)(5), but none of the exceptions would have applied on the alleged facts of the case. *See* 18 U.S.C. § 2702(b).

communications by the service provider.[181] The implicit threat model of the SCA is that outsiders, not the service provider itself, are the ones that might misuse the contents of communications. Thus, the law protects against both intrusions from the outside and disclosures to the outside, but not against misuse by insiders.[182] Such a threat model might have been sufficient in a world in which communications service providers did nothing but route communications. A communications service provider that is vertically integrated with other services, however, constitutes a new threat that lies outside the threat model of the SCA.

The lesson of *Councilman* is that a privacy law is only as strong as its threat model. It may well be that in a particular context, the law ought to ignore certain threats, but if so, it should be by design, rather than by oversight. Informed policy choices depend on appropriately identifying the relevant threats in a given context.

2. Characterizing Threats

Once we have identified a relevant threat, we then need to understand the nature of that threat. This encompasses both what harm a potential adversary might try to accomplish and what tools the adversary might use to accomplish that harm.

Defining the adversary's goal, or what counts as a privacy breach, has been one of the most important points of implicit disagreement among commentators and researchers writing about de-identification. Brickell and Shmatikov, for example, define a privacy breach in terms of "sensitive attribute disclosure."[183] In other words, their privacy goal is to hide some sensitive fact about a person from the adversary. As described in Part I, by relying on this study and others like it, Ohm implicitly adopts the same perspective.[184]

On the other hand, the El Emam meta-study is focused on record re-identification, that is, the ability of the adversary to determine the identity associated with a particular record in the data set.[185] Yakowitz also adopts this perspective.[186] So too

181. 18 U.S.C. § 2701(c)(1) (excluding conduct authorized "by the person or entity providing a wire or electronic communications service").

182. *See* 18 U.S.C. §§ 2701–2702.

183. Brickell & Shmatikov, *supra* note 64, at 70.

184. *See supra* Part I.

185. *See* El Emam et al., *supra* note 144, at 3.

186. *See supra* Part II.

do Schwartz and Solove, who propose applying different legal protections depending on the "risk of identification," where "identification" is defined to mean the "singl[ing] out [of] a specific individual from others."[187] As we have seen, identity disclosure and sensitive attribute disclosure are quite different conceptions of the adversary's goal because a data set can disclose sensitive attributes without also disclosing the identity associated with any particular record.[188]

The danger of not recognizing the distinction between different goals lies in implicitly adopting an underinclusive model that fails to capture relevant privacy harms. For example, by focusing only on identity disclosure, Schwartz and Solove miss the fact that the risk assessment they propose can be too narrow when the risk of sensitive attribute disclosure is high, but the risk of identity disclosure is low.[189] Moreover, their assumption that identity disclosure is the relevant risk masks important normative questions about how to define the nature of the risk rather than its magnitude. Schwartz and Solove cite literature on the factors that affect the risk of identity disclosure,[190] but those factors are of little help in deciding whether, for instance, to regard a prediction about a particular person's disease status as privacy-invading or socially useful.[191]

Apart from specifying the adversary's goals, we also need to specify the adversary's capabilities. One type of capability is the adversary's sophistication and computational power. One can reasonably assume that no adversary has unlimited processing power.[192] Beyond that, commentators debate

187. Schwartz & Solove, *supra* note 27, at 1877–78.

188. *See supra* note 135 and accompanying text.

189. *See* Schwartz & Solove, *supra* note 27, at 1879.

190. *See id.* (citing Khaled El Emam, *Risk-Based De-Identification of Health Data*, IEEE SECURITY & PRIVACY, May/June 2010, at 64); *see also* El Emam, *supra*, at 65 ("I focus on . . . *identity disclosure*.").

191. *See infra* Parts III.B–III.C.

192. *See* Ilya Mironov et al., *Computational Differential Privacy*, 5677 LECTURE NOTES IN COMPUTER SCIENCE (ADVANCES IN CRYPTOLOGY—CRYPTO 2009) 126 (2009). The technical term for this is that the adversary is "computationally-bounded." *See id.* The idea is not that the adversary is limited by the processing power of existing computers, but that there must be some outer limits to how many steps the adversary can perform, and that, as a result, there are certain "hard" problems that no conceivable adversary will ever be able to compute the answers to. This is the same assumption that underlies essentially all of modern data security, including, for example, secure transactions over the Internet. *See, e.g.*, *The Transport Layer Security (TLS) Protocol*, THE INTERNET ENGINEERING TASK FORCE (2008), *available at* http://tools.ietf.org/html/rfc5246.

whether to regard adversaries as mathematically sophisticated or not.[193] Again, any reasonable answer is surely contextual—marketers and identity thieves are presumably more sophisticated on the whole than the average person.

In assessing what sophistication the adversary needs, one should distinguish between the complexity of the science of re-identification and the complexity of the practice. The science might be complex, but an adversary may not need to know the science in order to carry out the re-identification. The actual techniques the adversary uses can be as simple as matching two sets of information.[194] Much depends on how much information the adversary has access to. It takes little sophistication to query a database and then dig around in the query results looking for additional matching background information. Anyone who has searched for a name on the Internet and tried to disambiguate the results has done this. Sophistication may well be necessary to assess whether an apparent match is likely to be an actual match,[195] but whether such an assessment is necessary to the adversary's goal is itself a contextual question. An identity thief who is risking being caught may want to be quite certain about the information he is using; a marketer can probably afford to just take a chance.

Background information is another resource available to the adversary. Commentators and researchers have also disagreed about whether and how to make assumptions about the adversary's background information.[196] Part of the difficulty in making such assumptions is that those assumptions can create a feedback loop. That is, if the law assumes the adversary knows relatively little, that assumption may provide the basis for justifying broader public disclosures of data. Those broad disclosures may in turn add to the adversary's knowledge in a way that breaks the assumptions that led to broad disclosures in the first place. Thus, it is

193. *See* Yakowitz, *supra* note 25, at 31–33.

194. *See supra* note 132 and accompanying text.

195. *See* Yakowitz, *supra* note 25, at 33 ("[D]esigning an attack algorithm that sufficiently matches multiple indirect identifiers across disparate sources of information, *and assesses the chance of a false match*, may require a good deal of sophistication.") (emphasis added).

196. *Compare* Ohm, *supra* note 21, at 1724 ("Computer scientists make one appropriately conservative assumption about outside information that regulators should adopt: We cannot predict the type and amount of outside information the adversary can access."), *with* Yakowitz, *supra* note 25, at 23 ("Not Every Piece of Information Can Be an Indirect Identifier").

important not only to characterize existing threats, but to assess how robust that characterization is to potential changes in the information environment.

3. Insiders and Outsiders

Another lesson of the *Councilman* case is that threats can differ as to whether they are "insider" or "outsider" threats. Privacy "insiders" are those whose relationship to a particular individual allows them to know significantly more about that individual than the general public does. Family and friends are examples. Co-workers might be insiders too. Service providers, both at the corporate and employee levels, could also be insiders, for example, employees at a communications service provider,[197] or workers at a health care facility.[198]

In security threat modeling, analysts regard insider attacks as "exceedingly difficult to counter," in part because of the "trust relationship . . . that genuine insiders have."[199] In the arena of data privacy, too, it can be similarly difficult to protect against disclosure to insiders, who can exploit special knowledge gained through their relationships with a target individual to deduce more about that individual from released data than the general public would. Protecting against privacy insiders may therefore require far greater restrictions on data release than protecting against outsiders.

Privacy law has never had a consistent answer to the question of whether the law targets only outsiders, or insiders as well. Consider the common law tort of public disclosure of private facts.[200] Traditionally, the rule has been that recovery under the tort requires a disclosure to the public at large, and not merely one that goes to a small number of individuals.[201]

---

197. *See, e.g.*, United States v. Councilman, 418 F.3d 67, 70 (1st Cir. 2005) (en banc).

198. *Cf.* Latanya Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, 25 J. L. MED. & ETHICS 98, 101 (1997) (describing the problem that "[n]urses, clerks and other hospital personnel will often remember unusual cases and, in interviews, may provide additional details that help identify the patient").

199. LANDAU, *supra* note 163, at 162–63.

200. *See* RESTATEMENT (SECOND) OF TORTS § 652D (1977).

201. *See, e.g.*, Wells v. Thomas, 569 F. Supp. 426, 437 (E.D. Pa. 1983) (finding "[p]ublication to the community of employees at staff meetings and discussions between defendants and other employees" insufficient to constitute "publicity"); Vogel v. W.T. Grant Co., 327 A.2d 133, 137 (Pa. 1974) (finding notification of "three relatives and one employer" insufficient to constitute "publicity"). In this

This is true even if the plaintiff was primarily trying to hide the information from a few people and only cared about what those few individuals knew.[202] Thus, one who discloses infidelity to a person's spouse is not liable, even though that may be the one person who matters.

The potential disconnect between a strict publicity requirement and what privacy plaintiffs actually care about, however, has led some courts to interpret the requirement in a more relaxed manner. Thus, in the case of *Beaumont v. Brown*, the court stated:

> An invasion of a plaintiff's right to privacy is important if it exposes private facts to a public whose knowledge of those facts would be embarrassing to the plaintiff. Such a public might be the general public, if the person were a public figure, or a particular public such as fellow employees, club members, church members, family, or neighbors, if the person were not a public figure.[203]

In other words, for private figures at least, disclosure to insiders such as "fellow employees, club members, church members, family, or neighbors" might suffice to make out a privacy tort claim.[204]

Similarly, identifiability with respect to insiders may be enough for a statement to be considered "of or concerning the plaintiff" for purposes of defamation or privacy law. In *Haynes v. Alfred A. Knopf, Inc.*, Judge Posner rejected the idea that the defendant should have redacted the names of the plaintiffs, finding that insiders would have been able to identify them anyway:

---

way, the requirement of "publicity" for a privacy tort is distinct from the element of "publication" for purposes of a defamation claim. A defamatory publication occurs when the statement is transmitted to any third party. *See* RESTATEMENT (SECOND) OF TORTS § 577 (1977).

202. *See Wells*, 569 F. Supp. at 437 ("Plaintiff's assertion that disclosures to the employees constituted publication to 'almost the entire universe of those who might have some awareness or interests in such facts,' even if assumed to be true, would not constitute 'publicity' but a mere spreading of the word by interested persons in the same way rumors are spread."). *Cf.* Sipple v. Chronicle Publ'g Co., 201 Cal. Rptr. 665, 667, 669 (Cal. App. 1984) (finding that plaintiff's sexual orientation was not a private fact, because it was "known by hundreds of people in a variety of cities," even though "his parents, brothers and sisters learned for the first time of his homosexual orientation" from the defendant).

203. Beaumont v. Brown, 257 N.W.2d 522, 531 (Mich. 1977).

204. *Id.*

> [T]he use of pseudonyms would not have gotten Lemann and Knopf off the legal hook. The details of the Hayneses' lives recounted in the book would identify them unmistakably to anyone who has known the Hayneses well for a long time (members of their families, for example), or who knew them before they got married; and no more is required for liability either in defamation law . . . or in privacy law.[205]

On the other hand, existing regulatory regimes largely ignore insiders with specialized knowledge.[206] The HIPAA safe harbor, for example, defines de-identified data to include any data with a specific list of eighteen identifiers removed.[207] The implicit threat model of such a safe harbor is one in which adversaries might know these particular identifiers, but no others. Even so, the HIPAA safe harbor contains the caveat that the entity releasing the data must "not have actual knowledge that the information *could be used alone or in combination with other information* to identify an individual who is a subject of the information."[208] Such language at least keeps open the possibility of including insiders in the threat model.

Whether to account for insiders is a question that must ultimately be resolved in context. For example, in *Northwestern Memorial Hospital v. Ashcroft*, the Seventh Circuit affirmed the district court's order quashing a government subpoena for redacted hospital records of women who had undergone late-term abortions.[209] Writing for the majority, Judge Posner held that redacting identity information was not enough to protect these women's privacy because of the significant risk that "persons of their acquaintance, or skillful 'Googlers,' sifting the information contained in the medical records concerning each patient's medical and sex history, will put two and two together, 'out' the 45 women, and thereby expose them to threats, humiliation, and obloquy."[210] Judge Posner's concern was, at least in part, about the potential for a breach by

205. Haynes v. Alfred A. Knopf, Inc., 8 F.3d 1222, 1233 (7th Cir. 1993) (citations omitted).

206. *See* Yakowitz, *supra* note 25, at 24–25.

207. *See* 45 C.F.R. § 164.514(b)(2) (2012).

208. *Id.* § 164.514(b)(2)(ii) (emphasis added).

209. 362 F.3d 923, 939 (7th Cir. 2004).

210. *Id.* at 929.

insiders. But as he noted, this was "hardly a typical case in which medical records get drawn into a lawsuit."[211] Rather, the records were part of a "long-running controversy over the morality and legality of abortion," in which there were "fierce emotions" and "enormous publicity."[212] When the privacy stakes are high, it may well be sensible to adopt a broader threat model, one that protects against "acquaintances" and other insiders, as well as against outsiders whose knowledge is derived only from Google searches.

4. Addressing Threats

After identifying and characterizing the relevant privacy threats arises the more normative question of which threats to address and which to ignore. Are concrete harms like discrimination or fraud the most appropriate threats to address? Should we address emotional harms that result when others think ill of us?[213] Or should we address the potential chilling effect of knowing that we may be subject to scrutiny?[214] Imagine a complete, searchable medical records database in which standard demographic information cannot be used to identify a record, but in which additional information, such as the date of a specific medical visit, can. Should we care that friends and family might be able to use such a database to discover our full medical records based on their knowledge of a few medical incidents?

Clearly, these fundamental questions about the nature of privacy cannot be settled here. The important point is that one's conception of privacy defines the universe of threats worth addressing, which in turn defines what it means to ensure "privacy" in released data. For example, to the extent our conception of privacy encompasses the more psychic and emotional harms that tend to result from revealing our secrets to acquaintances, rather than to strangers, we may be more inclined to regard revelations to those with significant non-public knowledge as something we ought to try to prevent.[215]

---

211. *Id.*

212. *Id.*

213. *See* SOLOVE, *supra* note 20, at 175–76.

214. *See* M. Ryan Calo, *The Boundaries of Privacy Harm*, 86 IND. L.J. 1131, 1145–47 (2011).

215. Psychic harm could be a component of revelations to strangers too in certain circumstances. *Cf.* Nw. Mem'l Hosp. v. Ashcroft, 362 F.3d 923, 929 (7th Cir. 2004) ("Imagine if nude pictures of a woman, uploaded to the Internet

Beyond the question of *which* threats to address lies the question of *how* to address them. In particular, law and technology are each tools that policymakers can use to mitigate threats, and each may be more appropriate or effective with respect to different types of threats.

In the security realm, one can characterize the anti-circumvention provisions of the Digital Millennium Copyright Act ("DMCA") as having adopted such a mixed strategy.[216] The DMCA imposes liability on one who "circumvent[s] a technological measure that effectively controls access to a [copyrighted] work."[217] The "technological measure that effectively controls access" prevents unauthorized access by the casual user, while liability under the DMCA itself addresses access by those with the technical sophistication to circumvent the system.[218] Technology addresses one set of threats, while the law fills in the gaps left by the technology.

The context of privacy-preserving data release may warrant a similar approach, with the form of the data addressing some threats, while law or regulation addresses others.[219] In particular, because insider threats are more difficult to address through technological means, legal solutions might be more appropriate for these threats. Similarly, legal controls might be particularly appropriate for more sophisticated threats.

The FTC's approach to defining the scope of its consumer privacy framework can be understood in this light. That framework applies to "data that can be reasonably linked to a specific consumer, computer, or other device."[220] In determining what data sets fall outside this definition, the FTC first requires that the data set be "not reasonably

---

without her consent though without identifying her by name, were downloaded in a foreign country by people who will never meet her. She would still feel that her privacy had been invaded.").

216. *See* 17 U.S.C. §§ 1201–1205 (2012). In analyzing the structure of the DMCA, I make no claim about its wisdom, which is beyond the scope of this Article.

217. *Id.* § 1201(a)(1)(A).

218. *See* Universal City Studios, Inc. v. Reimerdes, 111 F. Supp. 2d 294, 317–18 (S.D.N.Y. 2000) (holding that even a technological measure based on a "weak cipher" does "effectively control access" within the meaning of the statute).

219. *Cf.* Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 33, 47 (2010) (proposing "a statutory framework that will allow the data disclosers and the data recipients to agree voluntarily on externally enforceable terms that provide privacy protections for the data subjects").

220. FED. TRADE COMM'N, *supra* note 35, at 22.

identifiable."[221] Such a requirement perhaps ensures that the casual, rogue employee is not able to find juicy tidbits in the data set. As a whole, however, the company holding the data presumably has the sophistication and resources, as well as the inside knowledge, to circumvent more readily whatever mathematical transformations it applied to the data. Thus, the FTC also requires that the company itself "publicly commit[] not to re-identify" the data set, and that it similarly bind "downstream users" of the data.[222] As with the DMCA, in the FTC's framework, technology addresses one set of threats, and law addresses others.

Interpreting the FTC document in this way exposes ambiguities in the proposal, as well as how a threat modeling approach might help to resolve those ambiguities. It is not clear when a data set has been sufficiently transformed such that it is no longer "reasonably identifiable" under the FTC framework. Moreover, there is ambiguity as to what actions on the part of the company would constitute "re-identifying" the data. In both cases, those ambiguities should be resolved by determining what threats either the technology on the one hand, or the law on the other, are meant to address. For example, if an online advertising company uses the data to create a targeting program that is so fine-grained that it effectively personalizes advertising to each individual, has it "re-identified" the data? It may be difficult to derive any information about individuals by simply inspecting the targeting program itself, but if the ultimate harm we seek to prevent is the targeting of the advertisements, rather than the form in which the data is maintained, such a targeting program perhaps ought to be considered re-identification. Focusing on identifying and characterizing the relevant threats helps to give content to the legal standards intended to address those threats.

B. Uncertain Information

An important aspect of characterizing privacy threats is determining how to treat an adversary's acquisition of partial, or uncertain, information. Suppose, for instance, an adversary is 50 percent sure that a particular person has a particular

221. *Id.*
222. *Id.*

disease, or that a particular record belongs to a particular person. Different researchers have adopted very different assumptions in this respect. Brickell and Shmatikov count as a privacy loss any reduction in uncertainty about a subject's sensitive information.[223] El Emam, on the other hand, only counts verified identifications of individual records in the database.[224] Focusing on the relevant threats is key to assessing the significance of uncertain information.

A natural first instinct is to assume that uncertain information represents a risk of harm, so that 50 percent certainty about a person's disease status is equivalent to a 50 percent risk that the person's sensitive information will be disclosed. Following this instinct would lead one to approach the privacy question by looking to how the law generally treats a risk of harm, such as a 50 percent chance that a person will develop a disease.

The problem of risk of harm has been addressed within tort law under the rubric of the "loss of chance" doctrine.[225] This doctrine originated in the context of medical malpractice cases in which the doctor's negligence deprived the plaintiff of some chance of survival, such as through failure to diagnose cancer at an early stage.[226] Under the traditional rules of causation, if the patient died but did not have better than even odds of survival even with the correct diagnosis, then the courts denied recovery under the theory that it was more likely than not that the doctor's negligence made no difference in the end.[227] The loss of chance doctrine evolved out of a sense that the traditional doctrine was both unfair and resulted in under-deterrence.[228] Under a loss of chance theory, the relevant harm or injury is not simply the ultimate death or other medical injury, but rather the deprivation of "a chance to survive, to be cured, or otherwise to achieve a more favorable medical outcome," and the plaintiff can recover for the loss of that

223. *See* Brickell & Shmatikov, *supra* note 64, at 71–72.

224. *See* El Emam et al., *supra* note 144, at 3.

225. *See* Matsuyama v. Birnbaum, 890 N.E.2d 819, 823 (Mass. 2008). *See generally* David A. Fischer, *Tort Recovery for Loss of a Chance*, 36 WAKE FOREST L. REV. 605 (2001); Joseph H. King, Jr., *Causation, Valuation, and Chance in Personal Injury Torts Involving Preexisting Conditions and Future Consequences*, 90 YALE L.J. 1353 (1981).

226. *See* Matsuyama, 890 N.E.2d at 825–26.

227. *See id.* at 829.

228. *Id.* at 830.

chance.[229]

Some scholars have advocated that the loss of chance principle ought to apply equally to all cases in which the defendant's negligence increases the plaintiff's risk of future harm, even if that harm has not yet materialized.[230] Courts have been reluctant though to allow recovery for the risk of future harms, at least beyond the medical malpractice context.[231] In toxic tort cases, for example, several courts have not allowed plaintiffs to recover directly for the future risk of developing cancer or other diseases when such diseases are not reasonably certain to occur.[232] On the other hand, some courts have allowed plaintiffs to recover for other types of present injuries that flow from, but are not identical to, the risk of future harm, such as medical monitoring costs,[233] or emotional distress.[234]

One might view privacy harms through the lens of such tort cases, and indeed, such an analogy has already been made in the context of data breach litigation.[235] In data breach cases, courts have tended to reject even recovery for credit monitoring costs and emotional distress, let alone the pure risk of identity

229. *Id.* at 832.

230. *See* Ariel Porat & Alex Stein, *Liability for Future Harm*, in PERSPECTIVES ON CAUSATION 234–38 (Richard S. Goldberg, ed., 2010).

231. *See, e.g.*, Dillon v. Evanston Hospital, 771 N.E.2d 357, 367 (Ill. 2002) (describing as the "majority view" that "recovery of damages based on future consequences may be had only if such consequences are 'reasonably certain,'" where "reasonably certain" means "that it is more likely than not (a greater than 50 [percent] chance) that the projected consequence will occur"); *see also* Matsuyama, 890 N.E. at 834 n.33 (expressly limiting its decision to "loss of chance in medical malpractice actions" and reserving the question of "whether a plaintiff may recover on a loss of chance theory when the ultimate harm (such as death) has not yet come to pass"). The court in *Dillon* went on to reject the traditional rule, holding that the plaintiff could recover for the increased risk of future injuries caused by her doctor's negligence, even if such injuries were "not reasonably certain to occur." 771 N.E.2d at 370; *see also* Alexander v. Scheid, 726 N.E.2d 272 (Ind. 2000); Petriello v. Kalman, 576 A.2d 474 (Conn. 1990).

232. *See* Sterling v. Velsicol Chemical Corp., 855 F.2d 1188, 1204 (6th Cir. 1988); Ayers v. Jackson, 525 A.2d 287, 308 (N.J. 1987).

233. *See* Potter v. Firestone Tire & Rubber Co., 863 P.2d 795, 821–25 (Cal. 1993); *Ayers*, 525 A.2d at 312. *See generally* Andrew R. Klein, *Rethinking Medical Monitoring*, 64 BROOK. L. REV. 1 (1998).

234. *See* Eagle-Picher Indus., Inc. v. Cox, 481 So.2d 517 (Fla. Dist. Ct. App. 1985). *See generally* Andrew R. Klein, *Fear of Disease and the Puzzle of Futures Cases in Tort*, 35 U.C. DAVIS L. REV. 965 (2002).

235. *See* Vincent R. Johnson, *Credit-Monitoring Damages in Cybersecurity Tort Litigation*, 19 GEO. MASON L. REV. 113, 124–25 (2011) ("Data exposure and toxic exposure are analogous in that they both create a need for early detection of potentially emerging, threatened harm.").

theft or other data misuse.[236] In finding a lack of Article III standing, some courts have even questioned whether data spills cause *any* harms in the absence of misuse, and not just whether such harms are compensable.[237] If certainty of sensitive attribute disclosure or of identity disclosure is the relevant harm, then one might see support in the data breach cases for the view that actual re-identification, not mere "theoretical risk," should be the aim of any regulatory response.[238]

Uncertain information and risk of harm are not equivalent, however. Adversaries can have uncertain information without there being any significant risk of them obtaining the same information with certainty. Imagine a database in which ten records are precisely identical, except that five indicate a cancer diagnosis, while the other five indicate no cancer diagnosis. An adversary who is able to determine that a target individual must be one of these ten individuals can determine that there is a 50 percent chance that the person has cancer. However, because the ten records are otherwise identical, it is mathematically impossible for the adversary to use this data to determine the target individual's cancer status with certainty.[239]

---

236. *See* Pisciotta v. Old Nat'l Bancorp, 499 F.3d 629, 640 (7th Cir. 2007); Pinero v. Jackson Hewitt Tax Service Inc., 594 F. Supp. 2d 710, 715–16 (E.D. La. 2009).

237. *See* Reilly v. Ceridian Corp., 664 F.3d 38, 46 (3d Cir. 2011), *cert. denied*, 132 S. Ct. 2395 (2012). *But see* Krottner v. Starbucks Corp., 628 F.3d 1139, 1143 (9th Cir. 2010) (finding the plaintiffs' allegation of "a credible threat of real and immediate harm stemming from the theft of a laptop containing their unencrypted personal data" to be sufficient to meet "the injury-in-fact requirement for standing under Article III").

238. Yakowitz, *supra* note 25, at 20. Tort law, of course, might fail to provide a remedy not because the risk is deemed not to be a harm in itself, but for other administrability reasons. *Cf. Potter*, 863 P.2d at 811 (finding that it might well be "reasonable for a person who has ingested toxic substances to harbor a genuine and serious fear of cancer" even if the cancer has a low likelihood of occurring, but nevertheless holding, for "public policy reasons . . . , that emotional distress caused by the fear of a cancer that is not probable should generally not be compensable in a negligence action").

239. Of course, the adversary could guess randomly and be correct half of the time, but without a way to verify the guess, he would not know when he was correct and thus would still have no certainty. Studies looking for re-identification of individual records appear not to account for such random guessing, instead requiring certainty in order for the re-identification of a particular record to be deemed successful. For example, the Kwok and Lafky study, cited by both Yakowitz and El Emam, looked for records with "unique combinations of attribute values" in order to identify candidates for re-identification. Kwok & Lafky, *supra* note 143, at 5. Such a procedure would have excluded the records in the

More importantly, an adversary does not need to be certain in order to cause relevant privacy harms. That is, if harm is defined not by the disclosure of certain information, but rather by the ultimate uses to which an adversary puts that disclosed information, those harmful uses can arise without the adversary needing to be certain about the information itself. In that sense, when an adversary is 50 percent certain that a particular person has cancer, a present harm may have already occurred, rather than merely a risk of a future harm.

To see why this may be so, it is useful to consider the categories of privacy harm that Ryan Calo describes.[240] The first, which Calo describes as "subjective privacy harms," is defined by "the perception of unwanted observation, broadly defined."[241] For such a harm to exist, it is enough that the subject feels watched. It matters little what the watcher actually finds, or, for that matter, whether there really is a watcher at all.[242] For example, some find behavioral marketing to be harmful because it induces a "queasy" feeling of being watched.[243] In such a situation, it is not the use the adversary makes of its knowledge that matters, but the effect on the data subject of knowing that the adversary has such knowledge. The fact that an adversary's knowledge is uncertain may not diminish, and certainly does not eliminate, subjective privacy harms of this sort.

The other type of privacy harms are "objective privacy harms," which are "harms that are external to the victim and involve the forced or unanticipated use of personal information," resulting in an "adverse action."[244] Adverse actions can include consequences ranging from identity theft to negative judgments by others to marketing against the person's

---

hypothetical example above of ten records with nearly identical information. Similarly, in advocating *k*-anonymity as sufficient privacy protection, Yakowitz notes that the parameter *k* is usually set "between three and ten." Yakowitz, *supra* note 25, at 45. This obviously would not prevent the adversary from making similar random guesses as in the example above.

240. *See* Calo, *supra* note 214, at 1142–43.

241. *Id.* at 1144.

242. *Id.* at 1146–47.

243. *See* Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES, (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0 ("If we send someone a catalog and say, 'Congratulations on your first child!' and they've never told us they're pregnant, that's going to make some people uncomfortable . . . . [E]ven if you're following the law, you can do things where people get queasy.").

244. Calo, *supra* note 214, at 1148–50.

interests.[245] Identity theft may be a situation in which the target is only harmed if the thief's information is correct, but in many other contexts, uncertain information is more than sufficient to lead to objective harm. People frequently make judgments about others based on uncertain information. If there is stigma attached to a particular disease, for example, that stigma is likely to arise if acquaintances think that there is a significant chance that a particular person has that disease, even if they are not entirely sure. Similarly, marketers act with incomplete information. Advertisers target on the basis of their best guesses about the consumers they target.[246] If the targeting itself is the harm, that harm occurs equally no matter how certain the advertiser is about the characteristics of the targeted consumer.

Moreover, the significance of uncertain information cannot be evaluated numerically, and is instead highly contextual. The law tends to treat 51 percent as a magical number,[247] or to use some other generally applicable threshold of significance.[248] What matters with respect to privacy, however, is what effect uncertain information has, and the effect of a particular numerical level of certainty can vary widely across contexts. There is surely not a single threshold for determining when someone's guesses about another person's disease status will cause the target individual to be treated differently. The baseline rate for a sensitive characteristic matters (e.g., the prevalence of a disease in the general population), but while in some cases, we may care about the additive increase in certainty,[249] in others we may care about the multiplicative increase.[250] In the case of a relatively rare, but sensitive,

245. *See id.* at 1148, 1150–51.

246. *See, e.g.*, Julia Angwin, *The Web's New Gold Mine: Your Secrets*, WALL ST. J., July 31, 2010, at W1 (describing how advertising networks target advertising on the basis of "prediction[s]" and "estimates" of user characteristics, and using "probability algorithms").

247. *See* Matsuyama v. Birnbaum, 890 N.E.2d 819, 829 (Mass. 2008).

248. In the context of trademark litigation, for example, courts generally consider a showing of confusion among 15–25 percent of the relevant market enough to show "likelihood of confusion." *See, e.g.*, Thane Int'l, Inc. v. Trek Bicycle Corp., 305 F.3d 894, 903 (9th Cir. 2002) (finding that "a reasonable jury could conclude that a likelihood of confusion exists" based upon a survey "from which a reasonable jury *could* conclude that more than one quarter of those who encounter [the defendant's] ads will be confused").

249. *Cf.* Brickell & Shmatikov, *supra* note 64, at 76 (charting the absolute difference in percentage points between the knowledge of the adversary with and without identifiers in the database).

250. *Cf.* Andrew R. Klein, *A Model for Enhanced Risk Recovery in Tort*, 56

disease (e.g., HIV), it reveals almost nothing if an adversary is able to "guess" that some individual is HIV-negative. What we really care about is whether an adversary can correctly guess that an individual is HIV-positive, even though such guesses only increase the adversary's overall correctness by a fraction of one percent.

How we regard uncertain information may also relate to our assumptions about the adversary's background knowledge and, in general, the adversary's ability to leverage uncertain information. Should we worry about mass disclosure of medical records if we were assured that public demographic information could only be used by an adversary to identify ten possible records that might correspond to a particular individual?[251] While we might not worry about a mere 10 percent certainty in the abstract, such a scheme might nevertheless give us pause, because the information-rich nature of the disclosure could make it relatively easy for an adversary to use only a small amount of non-public information to narrow the set of possible records further from ten records down to a few possible records, or even down to an exact match. Thus, even if identify disclosure is the relevant harm, the risk of disclosure to insiders may be substantially higher than the same risk with respect to outsiders. And as previously described, if we focus on other harms, even 10 percent certainty might be enough to cause harm.

C. Social Utility

Just as commentators disagree about how to conceptualize "privacy," so too do they disagree about how to conceptualize "utility."[252] These disagreements are related, particularly with respect to statistical information, which Yakowitz suggests is socially useful rather than privacy-invading.[253] The difficulty is in separating the "good" statistical information from the "bad,"

WASH. & LEE L. REV. 1173, 1177 (1999) (arguing for recovery for enhanced risk when the plaintiff can prove that the toxic exposure doubled her risk of future disease).

251. This corresponds to a guarantee of 10-anonymity.

252. *See supra* Parts I–II.

253. *See* Yakowitz, *supra* note 25, at 29 ("Indeed, the definition of privacy breach used by Brickell and Shmatikov is a measure of the data's utility; if there are group differences between the values of the sensitive variables, . . . then the data is likely to be useful for exploring and understanding the causes of those differences.").

breast cancer rates in Marin County from block-level data on HIV-status, for example.

It cannot be that every inference that can be drawn from the data counts as socially useful, since anything we might call a privacy invasion is itself an inference drawn from the data. True, there is a sense in which any inference contributes to knowledge, but to find all knowledge equally deserving of protection would be to define utility in a way that necessarily clashes with privacy.[254] If utility is to be a useful concept, we need to distinguish among inferences, with some being those of legitimate researchers and others being those of privacy-invading adversaries.

Generalizability is one way of distinguishing "research" or information of "social value" from information that potentially invades privacy.[255] The HIPAA Privacy Rule defines "research" as "a systematic investigation . . . designed to develop or contribute to generalizable knowledge."[256] One can think of the newsworthiness test with respect to the tort of public disclosure as making a similar distinction in part, where courts have distinguished between newsworthy information "to which the public is entitled" and "a morbid and sensational prying into private lives for its own sake."[257] One way in which the disclosure might be not just for the sake of prying is if it contributes to knowledge about a wider class of people.[258]

Generalizability, however, is a social and contextual question, not purely a mathematical one. Imagine a scenario in which the adversary knows the target individual's age, race, and approximate weight, and is trying to determine whether that individual has diabetes. Suppose that the database to be released shows that in a national sample that does not include

254. *Cf.* Eugene Volokh, *Freedom of Speech and Information Privacy: The Troubling Implications of a Right to Stop People from Speaking About You*, 52 STAN. L. REV. 1049, 1050–51 (2000) (characterizing information privacy laws an inevitably problematic under the First Amendment because they create "a right to have the government stop you from speaking about me").

255. *See* Yakowitz, *supra* note 25, at 6 (defining "research" for purposes of her article to be "a methodical study designed to contribute to human knowledge by reaching verifiable and generalizable conclusions").

256. 45 C.F.R. § 164.501 (2013).

257. Virgil v. Time, Inc., 527 F.2d 1122, 1129 (9th Cir. 1975) (citing RESTATEMENT (SECOND) OF TORTS § 652D (Tentative Draft No. 21, 1975)).

258. *Cf.* Shulman v. Group W Productions, Inc., 955 P.2d 469, 488 (Cal. 1998) (finding the broadcast of the rescue and treatment of an accident victim to be of legitimate public interest "because it highlighted some of the challenges facing emergency workers dealing with serious accidents").

the target individual, 50 percent of individuals of that age, race, and weight have diabetes. The adversary might then naturally infer that there is a 50 percent chance that the target individual has diabetes.[259] Far from being information that we would want to suppress, information about the prevalence of disease within a particular demographic group is precisely the type of information that is worthy of study and dissemination.[260] In this example, the database has potentially revealed information about the target individual even though that individual does not appear in the database.[261] Thus, the only basis for the adversary's confidence in his inference is confidence that the research results are in fact generalizable and apply to similarly situated individuals not in the database.

On the other hand, if the target individual is in the released database, the adversary's inference that the individual is 50 percent likely to have diabetes might or might not be based on socially useful information.[262] One possibility, for example, is that the released database again shows that people of the target individual's age, race, and weight are 50 percent likely to have diabetes, and the database covers the entire country, or some similarly large population. In that case, the diabetes information from which the adversary was able to find out about the target individual would seem to be useful because it applies to a broad population. The same could be said if the database is a statistically sound sample of the broader population.

A different possibility, though, is that the adversary's inference is based on information about a small group that is neither interesting in itself nor representative of some larger group. For example, suppose the adversary knows the target

259. Yakowitz suggests that such an inference "is often inappropriate" because it involves "the use of aggregate statistics to judge or make a determination on an individual." Yakowitz, *supra* note 25, at 30. However, while such an inference might be socially (or legally) inappropriate in a particular context because of norms or laws against discrimination, the statistical inference itself will often be perfectly rational.

260. *See id.* at 28–29.

261. *Cf. supra* Part I.B (discussing differential privacy).

262. This discussion assumes that the adversary knows whether the individual is in the database. If not, then as explained above, *supra* note 130, we can switch our frame of reference to the population from which the database was drawn. For example, if there are only two people in the entire population that match the background information that the adversary has, and one of those people is shown in the database as having diabetes, then the adversary can again infer that there is at least a 50 percent chance that the target individual has diabetes.

individual's exact birth date, and that information allows the adversary to determine that the target individual's record must be one of ten records, of which five show the individual as having diabetes. The adversary will again be able to infer that there is a 50 percent chance that the target individual has diabetes. In this case, though, such an inference is unlikely to generalize. First, birth month and day were used to define the "demographic subgroup" in this case, and those characteristics are unlikely to have any medical significance.[263] Moreover, even a substantial deviation from the baseline rate of diabetes is probably not statistically significant, given the small size of the resulting subgroup. As a result, such an inference probably should not be regarded as useful, because the information revealed is nothing more than that of ten specific individuals, rather than that of a cognizable "subgroup." In each of these scenarios, the data revealed a 50 percent chance that the target individual has diabetes, but only some of these revelations were generalizable, and hence useful.

The concept of differential privacy may help to distinguish socially useful results from privacy-invading ones, but even with respect to differential privacy, the mathematical concept does not map perfectly onto the social one. Recall that a differentially private mechanism is designed to answer accurately only those questions that do not depend significantly on the presence or absence of one person in the data set.[264] Differential privacy can therefore distinguish between revealing the incidence of diabetes in a large demographic subgroup, and revealing the incidence in some small collection of individuals, because any one person will have a much smaller effect on the large group statistic than on the small group one. Differential privacy does not, however, take into account the social meaning of the attributes in the data set. In some instances, studying a small set of people might be quite legitimate, even though each individual has a strong effect on the research results—an example might be a study of those with a rare disease. Conversely, some studies of large populations might be regarded as illegitimate because of the particular subject of study. Perhaps some would regard trying to predict pregnancy on the basis of consumer purchases to be an illegitimate goal, even though the research result would be

263. *But see infra* note 269 and accompanying text.
264. *See supra* notes 114–116 and accompanying text.

generalizable and not dependent on any one individual.[265]

Similarly, social context is also the basis for deciding which fields can be completely suppressed without affecting utility. Consider the near universal requirement to strip names from a data set.[266] First or last name alone will, for most people, be far less uniquely identifying than many of the identifiers commonly left in the data set. Even the combination of first and last name is often not unique.[267] The requirement to strip names is not necessarily based on their uniqueness, but also their perceived lack of utility. We assume that we have much to gain, and little to lose, in dropping names.[268] The same might be said of other identifiers as well, such as exact birth dates.[269]

The concept of utility is thus highly contextual, and computer science cannot tell us what kind of utility we should want. Computer science can tell us, however, which kinds of utility tend to be more compatible with privacy, and which are less.

In general, uses of data can be categorized according to the type of inference that the researcher is trying to draw from the data.[270] One type might be how the frequency of a particular medical diagnosis varies by race. Another might be the best software program for using medical histories and demographics to predict whether someone has a particular medical

265. *See* Duhigg, *supra* note 243.

266. *See* Ohm, *supra* note 21, at 1713; Yakowitz, *supra* note 25, at 44–45.

267. There were, at one point, three people named "Felix Wu" in computer science departments in Northern California. *See Homepage of Felix F. Wu*, UNIV. OF CAL., BERKELEY, http://www.eecs.berkeley.edu/Faculty/Homepages/wu-f.html (last visited Mar. 25, 2013); *Homepage of Shyhtsun Felix Wu*, UNIV. OF CAL., DAVIS, http://www.cs.ucdavis.edu/~wu/ (last visited Mar. 25, 2013).

268. *But see generally* Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991 (2004) (documenting the effect of African-American sounding names on resumes on callback rates).

269. *But see* Joshua S. Gans & Andrew Leigh, *Born on the First of July: An (Un)natural Experiment in Birth Timing*, 93 J. PUBLIC ECON. 246, 247 (2009) (documenting a dramatic difference between the number of births in Australia on June 30, 2004 and July 1, 2004, corresponding to a $3000 government maternity payment, which applied to children born on or after July 1); Joshua S. Gans & Andrew Leigh, *What Explains the Fall in Weekend Births?*, MELBOURNE BUS. SCH. (Sept. 26, 2008), http://www.mbs.edu/home/jgans/papers/Weekend%20Shifting-08-09-26%20(ms%20only).pdf (documenting that proportionately fewer births occur on the weekends and correlating the overall drop in weekend births to the rise in caesarian section and induction rates).

270. These are called "concept classes" in the literature. *See* Blum et al., *supra* note 118, at 610.

condition.[271] In the latter case, rather than starting with some hypothesis, such as that race affects a particular disease, the researcher is effectively trying to derive the hypothesis from the data itself.

Intuitively, inferring a hypothesis is potentially much more complex than testing one. Computer scientists have formalized this idea with a mathematical way to measure the complexity of a set of potential inferences.[272] Broadly speaking, concrete, easy-to-state hypotheses are far less complex than hypotheses that cannot be succinctly represented, and testing straightforward hypotheses while still preserving privacy is significantly easier than inferring hypotheses from a broader, more complex concept class.[273] Thus, looking for "evidence of discrimination or disparate resource allocation" in school testing data[274] may well be possible in a privacy-preserving manner because these tasks only require the researcher to ask relatively simpler questions of the data.

In contrast, consider the Netflix Prize contest, in which the goal was to build an algorithm that could better predict people's movie preferences. Such a goal is easily stated, but what was "learned" in the end is not. The algorithm that the winners of the contest wrote is complicated and certainly cannot be described in a few lines of text.[275] The universe of possible learning algorithms that could have been applied to the Netflix Prize is immense. When we are trying to preserve the behavior of this enormous, difficult-to-characterize class of

271. These are "classifiers." *See supra* notes 93–94 and accompanying text.

272. This quantity is known as the Vapnik-Chervonenkis, or VC, Dimension. Roughly speaking, the VC-dimension measures the ability of a class of inferences to fit arbitrary data. *See* MICHAEL J. KEARNS & UMESH V. VAZIRANI, AN INTRODUCTION TO COMPUTATIONAL LEARNING THEORY 50–51 (1994). The more data that can be fit by a class of inferences, the higher the VC-dimension. For example, consider the class of threshold functions, which are functions whose result depends only on whether a given quantity is above or below some threshold. A researcher might use such functions to determine whether a disease correlates with having more than a certain amount of some substance in the patient's blood, for example. Any two data points can be explained with an appropriate threshold function, but with three data points, if the one in the middle is different from the other two, then the data cannot be explained using a threshold function. The VC-dimension of threshold functions is therefore 2. *See id.* at 52.

273. *See* Blum et al., *supra* note 118, at 611 ("It is possible to privately release a dataset that is simultaneously useful for any function in a concept class of polynomial VC-dimension.").

274. *See* Yakowitz, *supra* note 25, at 17 (discussing the potential beneficial uses of the data requested in *Fish v. Dallas Indep. Sch. Dist.*, 170 S.W.3d 226 (Tex. App. 2005)).

275. *See* Narayanan & Shmatikov, *supra* note 8, at 124 n.9.

algorithms, the utility of the data for these purposes is much more fragile and much less compatible with privacy-preserving techniques.[276] Thus, privacy and utility will seem more at odds when commentators focus on tasks like data mining as the relevant form of utility than when they focus on statistical studies.

Strands of this distinction between types of utility can be found in the common law. Consider the common law's treatment of whether the disclosure of identifying information is newsworthy. In some cases, such as *Barber v. Time, Inc.*, courts have found that even though the overall subject matter was newsworthy, the disclosure of the plaintiff's identity was not.[277] In *Barber*, the plaintiff suffered from a rare disorder that was the subject of a magazine article, which included her name and photograph.[278] In affirming a jury verdict in the plaintiff's favor, the court found the identity information added little or nothing to the medical facts, which could have been easily presented without it.[279] The utility, here newsworthiness, lay only in those straightforwardly articulable medical facts.

In contrast, in *Haynes v. Alfred A. Knopf, Inc.*, Judge Posner had a very different view of the value of data.[280] In that case, the plaintiff objected to his past being recounted in the context of "a highly praised, best-selling book of social and political history" about the Great Migration of African-Americans in the mid-20th century.[281] The plaintiff was not a significant historical figure; he was just one of many.[282] And, as one of many, so he argued, there was no reason to use his name or the details of his life.[283] Judge Posner disagreed, saying that if the author had altered the story, "he would no longer have been writing history. He would have been writing fiction. The nonquantitative study of living persons would be abolished as a category of scholarship, to be replaced by the sociological

276. *See id.* at 124.
277. *See* 159 S.W.2d 291, 295 (Mo. 1942).
278. *See id.* at 293.
279. *See id.* at 295 ("It was not necessary to state plaintiff's name in order to give medical information to the public as to the symptoms, nature, causes or results of her ailment. . . . Certainly plaintiff's picture conveyed no medical information.").
280. 8 F.3d 1222 (7th Cir. 1993).
281. *Id.* at 1224.
282. *Id.* at 1233.
283. *Id.*

novel."[284] According to Judge Posner, "the public needs the information conveyed by the book, including the information about Luther and Dorothy Haynes, in order to evaluate the profound social and political questions that the book raises."[285] In other words, there was utility to the story not captured by a bare presentation of historical facts or by a "sociological novel." The public would learn something legitimate, something generalizable, but in doing so, it was virtually impossible to protect the plaintiff's anonymity.

Data mining has much in common with historical accounts as described by Judge Posner. In each case, because it is hard to specify precisely what the researcher or reader is trying to learn, it is hard to modify the data in a way that is sure to preserve its value for the researcher or reader. As with the historical account, much hinges on whether we include complex data mining and similar tasks within our conception of utility. If we do, then it may be harder to protect privacy through mathematical privacy-preserving techniques.

D. Unpredictable Uses

Beyond the problem of determining what types of data uses ought to count as socially useful, there is an additional problem of determining at the time of data release what future uses of the data we want to support. As we have seen, utility is not a property of data in the abstract, but a property of data in context. The trouble is that we often do not know precisely what that context will turn out to be.[286] If we knew ahead of time exactly what data uses we would want to support, we could then eliminate everything else. In an extreme case, the data administrator could simply publish the research result itself, rather than any form of the database. In reality, however, we do not know how data will be used, and we want to support multiple uses simultaneously.[287]

---

284. *Id.*
285. *Id.*
286. *See* Yakowitz, *supra* note 25, at 10–13.
287. *See* Brickell & Shmatikov, *supra* note 64, at 74 ("The unknown workload is an essential premise—if the workloads were known in advance, the data publisher could simply execute them on the original data and publish just the results instead of releasing a sanitized version of the data."); *see also* Narayanan & Shmatikov, *supra* note 8, at 124 ("[I]n scenarios such as the Netflix Prize, the purpose of the data release is precisely to foster computations on the data that have not even been foreseen at the time of release.").

On the other hand, it is impossible to support *all* possible future uses without giving up on privacy entirely. This is one of the lessons of the principle that the greater the complexity of the uses we want to support, the less privacy we can maintain.[288] Recall that even throwing away something as seemingly useless as names can affect utility.[289]

The problem of unpredictable uses is particularly important with respect to any proposed principle of data minimization or use limitation. Both of these principles are part of the Fair Information Practice Principles, which are sometimes used to define a set of privacy interests.[290] Data minimization provides that "organizations should only collect PII ("Personally Identifiable Information") that is directly relevant and necessary to accomplish the specified purpose(s) and only retain PII for as long as is necessary to fulfill the specified purpose(s)."[291] Use limitation provides that "organizations should use PII solely for the purpose(s) specified in the notice."[292] By assuming that foreseen purposes control the collection, use, and retention of data, both of these principles foreclose unexpected uses. Whether they are appropriate thus depends on whether the context is one in which unexpected uses play an important part in defining utility.

As this Part has shown, bare invocations of the concepts of "privacy" and "utility" hide several dimensions along which commentators have disagreed. Conceptualizing privacy requires us to identify and characterize the relevant privacy threats, which then provides a basis for determining whether and how to address those threats. Moreover, thinking in terms of threats highlights the extent to which threats materialize on the basis of uncertain information. Similarly, conceptualizing utility requires us to evaluate the social significance of information in context and to determine at the outset what types of inferences to support in released data. This framework will help policymakers to sort through competing claims about the effects of data release or of de-identification techniques and

288. *See supra* Part III.C.

289. *See supra* note 268 and accompanying text.

290. *See, e.g.*, *National Strategy for Trusted Identities in Cyberspace*, THE WHITE HOUSE 45 (Apr. 2011); *see also* Schwartz & Solove, *supra* note 27, at 1879–80.

291. *National Strategy for Trusted Identities in Cyberspace*, *supra* note 290, at 45.

292. *Id.*

to see more clearly the policy implications of different data regulations.

IV. TWO EXAMPLES

The framework developed above sheds light on a number of specific issues, including two that will be discussed here: privacy interests in consumer data and the value of broader dissemination of court records.

A. Privacy of Consumer Data

The use of consumer data for targeted marketing poses a challenge to privacy laws centered around personally identifiable information, because the specific identity of the person targeted may not be all that relevant to either the use that the marketer wants to make of the information or to the nature of any harm that the person may suffer.[293] In the framework developed here, the re-identification of specific records is not by itself the relevant threat.

Understanding the relevant threat is the key to understanding cases like *Pineda v. Williams-Sonoma Stores, Inc.* and *Tyler v. Michaels Stores, Inc.*, each of which held that a zip code can be "personal identification information."[294] In both cases, the defendants argued that a zip code covers too many people to be identifiable information as to any one of them.[295] Given this fact, it would be "preposterous" to treat zip codes alone as personally identifiable information in all contexts.[296]

But that is not what either court did. Each court held that a zip code alone could be personal information in the context of the specific statute at issue, and, even more precisely, in the context of the specific threats at which each statute was aimed. In *Pineda*, the court held that the relevant threat was that of companies collecting "information unnecessary to the sales transaction" for later use in marketing or other "business purposes."[297] Because information like a zip code could be used

293. *See* Schwartz & Solove, *supra* note 27, at 1848 (discussing the "surprising irrelevance of PII" to behavioral marketing).

294. Pineda v. Williams-Sonoma Stores, Inc., 246 P.3d 612, 614 (Cal. 2011); Tyler v. Michaels Stores, Inc., 840 F. Supp. 2d 438, 446 (D. Mass. 2012).

295. *Pineda*, 246 P.3d at 617; *Tyler*, 840 F. Supp. 2d at 442.

296. *See* Yakowitz, *supra* note 25, at 55 n.265.

297. 246 P.3d at 617.

to help "locate the cardholder's complete address or telephone number," excluding it from the statute "would vitiate the statute's effectiveness."[298] In contrast, in *Tyler*, the court held that the statute was aimed at the threat of "identity theft and identity fraud," not marketing.[299] Nevertheless, the result was the same because "in some circumstances the credit card issuer may require the [zip] code to authorize a transfer of funds," and, thus, the zip code could be "used fraudulently to assume the identity of the card holder."[300] In each case, zip codes were important to the threat model, but for entirely different reasons. It was a key piece of information that the companies collecting it could themselves use to link individual sales transactions to full addresses and marketing profiles.[301] It was also a key piece of information that, when written down, identity thieves might acquire and use to commit fraud.[302]

The end results in *Pineda* and *Tyler* aligned, but in general, the implications of focusing on the threat of marketing will be very different from the implications of focusing on the threat of identity theft. Much ordinary consumer transaction data may contribute to the effectiveness of targeted marketing,[303] but is unlikely to be particularly useful for identity theft. Thus, an important question for determining the appropriate scope of consumer data privacy laws is whether the marketing activity itself should be regarded as a relevant threat, or whether the threats are primarily those of unwanted disclosure or of fraudulent use of the information by outsiders. Privacy laws that treat the marketing itself as a relevant harm will be much broader than those aimed only at disclosure and fraud.

B. Utility of Court Records

Court records have long been regarded as public documents, but the greater ease with which access is now possible, as records become increasingly electronic and remotely available, has raised privacy concerns.[304] On the one

298. *Id.* at 618.
299. 840 F. Supp. 2d at 445.
300. *Id.* at 446.
301. *See Pineda*, 246 P.3d at 617.
302. *See Tyler*, 840 F. Supp. 2d at 446.
303. *See* Angwin, *supra* note 246.
304. *See* Amanda Conley et al., *Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry*, 71 MD. L. REV.

hand, much sensitive information is available in court records, ranging from social security numbers to sensitive medical facts, but on the other hand, there are important public functions to open court records that must be balanced against any privacy concerns. In the framework developed here, we must specify what utility we are seeking to obtain from the data.

One possibility is that court records, like all large compilations of rich social data, are an important source of sociological research.[305] As we have seen, whether such research can be supported in a privacy-protecting manner may depend on what "research" we have in mind.[306] Looking for specific types of patterns in the data may be easier to support than being able to mine the data for arbitrary and unpredictable patterns. Being able to gather statistical information is far easier to do privately than being able to use the data to tell a story.[307]

The interest most often asserted with respect to open court records is an interest in transparency and accountability.[308] Here too, it is necessary to specify more precisely what we mean by accountability. On one view, accountability may be an aggregate property, a feature of the workings of government as a whole. In that case, we may be able to achieve accountability and privacy at the same time by redacting, sampling, and modifying the released data. On a different view, however, accountability requires the government to be accountable in each individual instance. If it is not just that society deserves to see how the government as a whole is doing, but rather that each individual has a right to ensure that the government is doing right by every individual, then there is a more fundamental conflict between the accountability and privacy interests at stake. In this way, conceptions of accountability, a form of utility relevant here, are crucial to understanding the balance between privacy and utility with respect to access to court records.

772, 774 (2012).

305. *See* David Robinson et al., *Government Data and the Invisible Hand*, 11 YALE J.L. & TECH. 160, 166 (2009).

306. *See supra* Part III.C.

307. *See supra* notes 280–285 and accompanying text.

308. *See* Conley et al., *supra* note 304, at 836; *see also* Grayson Barber, *Personal Information in Government Records: Protecting the Public Interest in Privacy*, 25 ST. LOUIS U. PUB. L. REV. 63, 93 (2006) ("The presumption of public access to court records allows the citizenry to monitor the functioning of our courts, thereby insuring quality, honesty, and respect for our legal system.").

CONCLUSION

Although all sides in the debate over data disclosure hold up concepts and results from computer science to support their views, there is a more fundamental underlying debate, masked by the technical content. It is a debate about what values privacy ultimately serves. At the root of distrust of anonymization is a broad conception of "privacy" that includes protecting us from the guesses that our friends and neighbors might make about us. At the root of faith in anonymization is a significantly narrower conception of "privacy" that looks for more concrete harms like identity theft. Moreover, commentators implicitly disagree about what we ought to be able to do with data, whether more foreseeable statistical tasks or arbitrary, unforeseen discoveries. We must grapple, in context, with these fundamental issues of conceptualizing privacy and utility in data sets before we can determine what combination of anonymization and law to use to balance privacy and utility in the future.