Harvard Journal of Law & Technology
Volume 25, Number 1 Fall 2011

TRAGEDY OF THE DATA COMMONS

*Jane Yakowitz**

TABLE OF CONTENTS

I. INTRODUCTION 2

II. FRUITS OF THE DATA COMMONS 5
A. Research Data 6
B. The Value of the Data Commons 8
C. Ex Ante Valuation Problems 10
D. The Importance of Broad Accessibility 13
E. Freedom of Information Act Requests: Privacy as an Evasion Technique 17

III. DOOMSDAY DETECTION: THE COMPUTER SCIENCE APPROACH 20
A. How Attack Algorithms Work 21
B. Erroneous Assertions 23
1. Not Every Piece of Information Can Be an Indirect Identifier 23
2. Group-Based Inferences Are Not Disclosures 28
3. A Data Release Can Be Useful and Safe at the Same Time 30
4. Re-Identifying Subjects in Anonymized Data Is Not Easy 31
5. De-Anonymized Public Data Is Not Valuable to Adversaries 33

IV. THE SKY IS NOT FALLING: THE REALISTIC RISKS OF PUBLIC DATA 35
A. Defective Anonymization 36
B. The Probability that Adversaries Exist 37
C. Scale of the Risk of Re-Identification in Comparison to Other Tolerated Risks 39

---

* Visiting Assistant Professor of Law, Brooklyn Law School; Yale Law School, J.D.; Yale College, B.S. The author is grateful for invaluable feedback from Jeremy Albright, Jonathan Askin, Miriam Baer, Derek Bambauer, Daniel Barth-Jones, Anita Bernstein, Frederic Bloom, Ryan Calo, Deven Desai, Robin Effron, Khaled El Emam, Marsha Garrison, Robert Gellman, Eric Goldman, Dan Hunter, Ted Janger, Margo Kaplan, Claire Kelly, Bailey Kuklin, Rebecca Kysar, Brian Lee, David S. Levine, Andrea M. Matwyshyn, Bill McGeveran, Helen Nissenbaum, Paul Ohm, Richard Sander, Liz Schneider, Paul Schwartz, Christopher Soghoian, Larry Solan, Berin Szoka, Nelson Tebbe, Adam Thierer, Marketa Trimble, Felix Wu, and Peter Yu. This article was generously supported by the Brooklyn Law School Dean's Summer Research Stipend Program.

V. A PROPOSAL IN THE STATE OF HIGHLY UNLIKELY RISK 42
A. Anonymizing Data 44
B. Safe Harbor for Anonymized Data 47
C. Criminal Penalties for Data Abuse 48
D. Objections 50
E. Improving the Status Quo 53

VI. CONCLUSION: THE TRAGEDY OF THE DATA COMMONS 61
A. Problems with the Property Model 62
B. The Data Subject as the Honorable Public Servant 66

I. INTRODUCTION

Over the past ten years, the debate over welfare reform has been transformed by Jeffrey Grogger and his coauthors. Grogger's data-driven research shows, among other things, that work requirements and time limits may have no effect on marriage or fertility rates.[1] In other words, welfare does not produce "welfare queens." More recently, Roland Fryer and Steven Levitt have discredited Herrnstein's theory that the test score gap between Caucasians and African Americans is the result of biological differences. Fryer and Levitt used longitudinal data to document for the first time that there are no differences in the cognitive skills of white and black nine-month-old babies, and that the gap that develops by elementary school is explained almost entirely by socio-economic and environmental factors.[2] And in 2001, John J. Donohue and Steven D. Levitt presented shocking evidence that the decline in crime rates during the 1990s, which had defied explanation for many years, was caused in large measure by the introduction of legalized abortion a generation earlier.[3]

These studies and many others have made invaluable contributions to public discourse and policy debates, and they would not have been possible without anonymized research data — what I call the "data commons." The data commons is comprised of the disparate and

1. JEFFREY GROGGER & LYNN A. KAROLY, WELFARE REFORM: EFFECTS OF A DECADE OF CHANGE 196–97 (2005). Grogger has also produced empirical evidence that welfare-to-work reforms did lead to increased wages and increased rates of non-dependence among the welfare recipients, but also had a negative impact on the academic performance of their adolescent children. Jeff Grogger & Charles Michalopoulos, *Welfare Dynamics Under Term Limits* (Nat'l Bureau of Econ. Research, Working Paper No. 7353, 1999); Jeffrey Grogger, Lynn A. Karoly & Jacob Alex Klerman, *Conflicting Benefits Trade-Offs in Welfare Reform*, RAND.ORG (2002), http://www.rand.org/publications/randreview/issues/rr-12-02/benefits.html.

2. Roland G. Fryer, Jr. & Steven D. Levitt, *Understanding the Black-White Test Score Gap in the First Two Years of School*, 86 REV. ECON. & STAT. 447, 447 (2004); Roland G. Fryer, Jr. & Steven D. Levitt, *Testing for Racial Differences in the Mental Ability of Young Children* (Nat'l Bureau of Econ. Research, Working Paper No. 12066, 2006).

3. John J. Donohue III & Steven D. Levitt, *The Impact of Legalized Abortion on Crime*, 116 Q.J. ECON. 379 (2001).

diffuse collections of data made broadly available to researchers with only minimal barriers to entry. We are all in the data commons; information from our tax returns, medical records, and standardized tests seed the pastures. We are protected from embarrassment and misuse by anonymization. But a confluence of events has motivated privacy experts to abandon their faith in data anonymization.

In his recent article, Paul Ohm brought the concerns of the computer science community to a wide audience of lawyers and policymakers. Ohm's argument is simple and superficially sound: As the amount of publicly available information on the Internet grows, so too does the chance that a malfeasor can reverse engineer a dataset that was once anonymized and expose sensitive information about one of the data subjects.[4] Privacy advocates, the media, and the Federal Trade Commission ("FTC") have accepted uncritically the notion that anonymization is impossible, and they advocate for the wholesale dismantling of the concept of anonymization.[5] In its place, privacy advocates recommend that research data should be regulated under the strong property and autonomy models of privacy favored by Lawrence Lessig, Jerry Kang, Paul Schwartz, and other scholars.[6]

Today, data privacy practices are shaped by some combination of ambiguous statutory directives, inconsistent case law, industry best practices, whim, and self-serving discretionary preferences. The time is ripe for the creation of uniform data privacy policies, and there is much to fix.[7] But proposals that inhibit the dissemination of research data dispose of an important public resource without reducing the pri-

---

4. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010).

5. *See id. See generally* FTC, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: A PROPOSED FRAMEWORK FOR BUSINESSES AND POLICYMAKERS (2010) [hereinafter FTC PRIVACY REPORT], *available at* http://www.ftc.gov/os/2010/12/101201privacyreport.pdf; Ryan Singel, *Netflix Spilled Your* Brokeback Mountain *Secret, Lawsuit Claims*, WIRED THREAT LEVEL (Dec. 17, 2009, 4:29 PM), http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit; Seth Schoen, *What Information is "Personally Identifiable"?*, ELECTRONIC FRONTIER FOUND. DEEPLINKS (Sept. 11, 2009, 10:43 PM), http://www.eff.org/deeplinks/2009/09/what-information-personally-identifiable; *Re-identification*, ELECTRONIC PRIVACY INFO. CENTER, http://epic.org/privacy/reidentification/ (last visited Dec. 21, 2011). Parties in several recent lawsuits have argued that there is no longer a tenable difference between anonymized information and personally identifiable information. *See, e.g.*, Complaint at 20, Gaos v. Google Inc., No. 10-CV-04809 (N.D. Cal. May 2, 2011); Complaint at 15, Doe v. Netflix, No. C09 05903 (N.D. Cal. Dec. 17, 2009) [hereinafter Doe Complaint]; Elinor Mills, *AOL Sued over Web Search Data Release*, CNET NEWS BLOGS (Sept. 25, 2006, 12:17 PM), http://news.cnet.com/8301-10784_3-6119218-7.html.

6. *See, e.g.*, LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE 142–63 (1999); Jerry Kang & Benedikt Buchner, *Privacy in Atlantis*, 18 HARV. J.L. & TECH. 229, 255 (2004); Paul M. Schwartz, *Property, Privacy, and Personal Data*, 117 HARV. L. REV. 2055, 2076, 2088–113 (2004).

7. Privacy law is on the mind of politicians and regulators and has entered what John Kingdon calls the proverbial "policy window." JOHN KINGDON, AGENDAS, ALTERNATIVES, AND PUBLIC POLICIES 165 (2d ed. 2002).

vacy risks that actually put us in peril. This Article argues that it is in fact the research data that is now in great need of protection. People have begun to defensively guard anonymized information about themselves. We are witnessing a modern example of a tragedy of the commons.[8] Each individual has an incentive to remove her data from the commons to avoid remote risks of re-identification. This way she gets the best of both worlds: her data is safe, and she also receives the indirect benefits of helpful health and policy research performed on the rest of the data left in the commons. However, the collective benefits derived from the data commons will rapidly degenerate if data subjects opt out to protect themselves.[9]

This Article challenges the dominant perception about the risks of research data by making three core claims. First, the social utility of the data commons is misunderstood and greatly undervalued by most privacy scholars. Public research data produces rich contributions to our collective pursuit of knowledge and justice. Second, the influential legal scholarship by Ohm and others misinterprets the computer science literature, and as a result, oversells the futility of anonymization, even with respect to theoretical risk. And third, the realistic risks posed by the data commons are negligible. So far, there have been no known occurrences of improper re-identification of a research dataset. Even the hypothetical risks are smaller than other information-based risks (from data spills or hacking, e.g.) that we routinely tolerate for convenience.

The Article proceeds as follows: Parts II, III, and IV perform a risk-utility calculus on the data commons, finding that the public data commons is tremendously valuable (Part II), that the theoretical risks of research data are exaggerated (Part III), and that the true risks posed by research data are nonexistent (Part IV). Together, Parts II

---

8. The tragedy of the commons model I explore here is not perfectly analogous to the "grazing commons" concept popularized by Garrett Hardin. Garrett Hardin, *The Tragedy of the Commons*, 162 SCIENCE 1243 (1968). In the grazing model, self-interested actors convert the communal benefits of the commons into private benefits for themselves. The gain from adding one more cow of their own is internalized, while the losses in the form of overgrazing are externalized and borne by the entire population. *Id.* In the data commons, the data subject depletes the commons by removing his data. The marginal detriment of his decision is externalized and shared across the entire population. Meanwhile, he enjoys the full value of the avoided risk of re-identification. Unlike the traditional commons examples, each actor is constrained in how much of the commons he is capable of depleting since he has but one line of data to remove. (The grazing and pollution examples that Hardin discusses anticipate actors who deposit multiple cows, or increasing amounts of pollution, into the commons). But the key point is intact: communal benefits are lost due to actions motivated by self-interest. Vaccination makes an even better comparison. *See infra* Part VI.

9. Fred Cate makes a similar argument in the context of consumer data used for credit reports. *See* Fred H. Cate, Data and Democracy, Herman B Wells Distinguished Lecture of the Institute and Society for Advanced Study (Sept. 21, 2001), *in* IND. UNIV., INST. FOR ADVANCED STUDY AND SOC'Y FOR ADVANCED STUDY, HERMAN B WELLS DISTINGUISHED LECTURE SERIES 1 (2001), *available at* https://scholarworks.iu.edu/dspace/bitstream/handle/2022/8508/IAS-WDLS-01.pdf.

through IV show that concerns over anonymized data have all the characteristics of a moral panic and are out of proportion to the actual threat posed by data dissemination.[10] In Part V, I put forward a bold proposal to redesign privacy policy such that public research data would be *easier* to disseminate. While data users who intentionally re-identify a subject in an anonymized dataset should be sanctioned heavily, agencies and firms that compile and produce the data in the commons should receive immunity from statutory or common law privacy claims so long as they undergo basic anonymization techniques. Part V also provides clear guidance for data producers operating under the current statutory regime. Part VI concludes with an appeal to the legal community to think and talk about research data differently. The bulk of privacy scholarship has had the deleterious effect of exacerbating public distrust in research data. Rather than encouraging the public to fervently guard their self-interest, scholars should build a sense of civic responsibility to pay their "information taxes" and participate in research datasets.

II. FRUITS OF THE DATA COMMONS

The benefits flowing from the data commons are indirect but bountiful. Thus far, the nascent technical literature on de-anonymization has virtually ignored the opportunity costs that would result from a drastic reduction in data sharing.[11] Legal scholars who write on the topic acknowledge the public interest in information, but they give that interest short shrift and describe it in abstract terms.[12]

---

10. For a discussion of "moral panics," see STANLEY COHEN, FOLK DEVILS AND MORAL PANICS (1972). Here, advocacy groups' demand for political action is driven by fears that privacy and anonymity as we know them are on the brink of ruin.

11. For example, the Netflix de-anonymization study, on which Ohm relies heavily, makes no effort to compare the risk of re-identification to the utility of the dataset. Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 PROC. 29TH IEEE SYMP. ON SECURITY & PRIVACY 111. The early work of Latanya Sweeney acknowledged a tradeoff between a dataset's utility and its theoretical re-identification risk, but the discussion of utility was abstract and very brief. Moreover, Sweeney's recent work pays no regard to the countervailing interests in data utility at all. *Compare* Latanya Sweeney, Computational Disclosure Control: A Primer on Data Privacy Protection (May 2001) (unpublished Ph.D. thesis, Massachusetts Institute of Technology), *available at* http://dspace.mit.edu/bitstream/handle/1721.1/8589/49279409.pdf, *with* Latanya Sweeney, *Patient Identifiability in Pharmaceutical Marketing Data* (Data Privacy Lab Working Paper 1015, 2011), *available at* http://dataprivacylab.org/projects/identifiability/pharma1.pdf. The statistical literature on disclosure risk generally recognizes the tension between the utility of data sharing and its concomitant risks but struggles to define best practices that can persist with increasing amounts of data accumulation. For a review of the state of the current computer science literature on the subject, see GEORGE T. DUNCAN ET AL., STATISTICAL CONFIDENTIALITY: PRINCIPLES AND PRACTICE (2011).

12. *See, e.g.*, PAUL M. SCHWARTZ, THE CTR. FOR INFO. POLICY LEADERSHIP, DATA PROTECTION LAW AND THE ETHICAL USE OF ANALYTICS 8 (2010), *available at* http://www.huntonfiles.com/files/webupload/CIPL_Ethical_Undperinnings_of_Analytics_Paper.pdf; Ohm, *supra* note 4, at 1708, 1714. *But see, e.g.*, Douglas J. Sylvester & Sharon

To strike the right balance between the public's interest in privacy and its interest in the data commons, we must have a more concrete understanding of the value gleaned from broadly accessible research data. In this Part, I define the data commons and explore its utility. I also discuss government agencies' pretextual use of privacy law to evade Freedom of Information Act ("FOIA") requests when disclosures could reveal something embarrassing to the government.

A. Research Data

This Article addresses datasets that are compiled and shared broadly for "research," by which I mean a methodical study designed to contribute to human knowledge by reaching verifiable and generalizable conclusions.[13] Although this is an expansive definition of "research," it importantly excludes analytic studies on the subject pool for the purpose of understanding the particular individuals in the pool, as opposed to understanding a general population.[14]

Public-use research datasets are usually subject to legal constraints that guard the privacy of the data subjects, and the largest producers of research data (including the U.S. Census Bureau and other federal agencies) use sophisticated anonymization techniques that go well beyond the minimum legal requirements.[15] Privacy laws in their various forms usually prohibit the release of personally identifiable

---

Lohr, *The Security of Our Secrets: A History of Privacy and Confidentiality in Law and Statistical Practice*, 83 DENV. U. L. REV. 147, 196–99 (2005); 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, 1122–24 (2000).

13. 45 C.F.R. § 164.501 (2010) (defining research as "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge").

14. A business entity might be very interested in what the particular individuals in its, or a competitor's, databases are like and inclined to purchase, regardless of whether their analytics can be generalized to describe human phenomena. Data researchers are naturally indifferent to information about any particular person because information about that person cannot be generalized to any class of persons. "Statistical data are unconcerned with individual identities. They are collected to answer questions such as 'how many?' or 'what proportion?', not 'who?'. The identities and records of co-operating (or non-cooperating [sic]) subjects should therefore be kept confidential, whether or not confidentiality has been explicitly pledged." *ISI Declaration on Professional Ethics*, INT'L STAT. INST. (Aug. 1985), http://isi-web.org/about/ethics1985; *see also* Sylvester & Lohr, *supra* note 12, at 185.

15. *See*, *e.g.*, *Confidentiality Statement*, U.S. CENSUS BUREAU, http://factfinder.census.gov/jsp/saff/SAFFInfo.jsp?_pageId=su5_confidentiality (last updated Mar. 17, 2009). These techniques include top-coding, data swapping, and the addition of random noise. *See* Jerome P. Reiter, *Estimating Risks of Identification Disclosure in Microdata*, 100 J. AM. STAT. ASS'N 1103, 1103 (2005). While these techniques increase privacy, they come at a cost to the utility of the data since the fuzzied data affects the results of statistical analyses. *See, e.g.*, A. F. Karr et al., *A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality*, 60 AM. STATISTICIAN 224, 224 (2006). Data archivists and social scientists conceive of privacy obligations differently from lawmakers and, not surprisingly, their approach is more nuanced.

information ("PII").[16] Information is personally identifiable if it can be traced to a specific individual.[17] Obviously, information that is tied to a direct identifier, such as name, address, or social security number, is personally identifiable. For example:

> Jane Yakowitz is actually a giant cockroach.

However, PII is not limited to information that directly identifies a subject. Included in its ambit are pieces of information that can be used in combination to indirectly link sensitive information to a particular person.

> A 31-year-old white female who works at Brooklyn Law School and lives in ZIP code 11215 is actually a giant cockroach.

Or:

> All 31-year-old females that live in ZIP code 11215 are actually giant cockroaches.

I will use the term "indirect identifiers" to mean pieces of information that can lead to the identity of a person through cross-reference to other public sources or through general knowledge.[18] "Non-identifiers," in contrast, cannot be traced to individuals without having special non-public information.

Paul Ohm has criticized U.S. privacy law for using static definitions of what constitutes PII,[19] but his description of the law is inaccu-

---

16. See discussion of the Family Education Rights and Privacy Act ("FERPA"), Health Insurance Portability and Accountability Act ("HIPAA"), and the Confidential Information Protection and Statistical Efficiency Act *infra* text accompanying notes 20–21.

17. For example, the HIPAA Standards for Privacy of Individually Identifiable Health Information (the "HIPAA Privacy Rule") define individually identifiable information as information that "identifies the individual" or information "[w]ith respect to which there is a reasonable basis to believe the information can be used to identify the individual." 45 C.F.R. § 160.103 (2010).

18. I borrow this term from the Department of Education's commentary on the final ruling of the 2008 revisions to the FERPA regulations. Family Educational Rights and Privacy, 73 Fed. Reg. 74,806, 74,831 (Dec. 9, 2008). Although some use other terminology such as "high risk variables," I prefer the term "indirect identifier" because it connotes that the information might be usable for tracing an identity without implying that it always and necessarily heightens the risk of re-identification to an unacceptable level. Latanya Sweeney, the computer scientist at Carnegie Mellon University who popularized the k-anonymity model for de-identifying data, uses the term "quasi-identifiers." Latanya Sweeney, *k-Anonymity: A Model for Protecting Privacy*, 10 INT'L J. UNCERTAINTY, FUZZINESS AND KNOWLEDGE-BASED SYSTEMS 557, 563 (2002).

19. Ohm, *supra* note 4, at 1740–41. Paul Ohm suggests modifying the rhetoric used in information privacy to connote that common privacy techniques merely "try to achieve anonymity," and do not actually achieve it. *Id.* at 1744. I like his recommendation to use the

rate. Privacy statutes list categories of information that necessarily must be classified as indirect identifiers (such as sex and ZIP code), but the statutes also obligate data producers to guard against other unspecified indirect identifiers that, in context, could be used to re-identify a subject. For example, the Confidential Information Protection and Statistical Efficiency Act ("CIPSEA") disallows the disclosure of statistical data or information that is in "identifiable form," defined as "any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means."[20] The Family Education Rights and Privacy Act ("FERPA") and the regulations implemented under the Health Insurance Portability and Accountability Act ("HIPAA") define PII similarly, with savings clauses that prohibit releases that might be reverse engineered through indirect means.[21]

The PII standard has a significant impact on the data commons. Large, information-rich datasets will inevitably contain PII because the combinations of indirect identifiers are likely to make some of the subjects unique, or close to it. Thus, even the legal minimum anonymization requires some of the utility of a dataset to be lost through redaction and blurring in order to ensure that no subject has a unique combination of indirect identifiers.

B. The Value of the Data Commons

In 1997, policy researchers at the RAND Corporation warned that the Sentencing Reform Act of 1984 and the plethora of state statutes setting minimum sentencing requirements for drug convictions are a less cost-effective means to reduce the consumption of cocaine than

---

term "scrub," *id.*, but Ohm's linguistic analysis reveals something about his assumptions. To Ohm, there never *was* a difference between *trying to achieve* anonymity and anonymity; anonymization techniques were never believed to be completely without risk.

20. E-Government Act of 2002, Pub. L. No. 107-347, § 502(4), 116 Stat. 2962, 2962 (codified at 44 U.S.C. § 3501 note).

21. *See* FERPA, 20 U.S.C.A. § 1232g (West 2010 & Supp. 2011); HIPAA Standards for Privacy of Individually Identifiable Health Information, 45 C.F.R. § 160.103 (2010). Usually, multiple indirect identifiers have to be combined in order to ascertain the identity of a specific individual. Privacy law is mindful of this potential route to re-identification and explicitly guards against it — any combination of publicly knowable information that can be used to trace to an identity is PII. The FERPA regulations prohibit the disclosure of "[o]ther information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty." 34 C.F.R. § 99.3 (2010). The HIPAA Privacy Rule prohibits the disclosure of "protected health information," 45 C.F.R. § 164.502 (2010), including information "(i) [t]hat identifies the individual; or (ii) [w]ith respect to which there is a reasonable basis to believe the information can be used to identify the individual." *Id.* § 160.103.

the previous system.[22] Moreover, both enforcement regimes were less effective per dollar spent on enforcement than on treatment programs.[23] While the change in policy could be defended on the basis of retributive goals, the promised deterrent effects were illusory.[24] Now that states are facing gaping budget holes, the tune has changed. The severity and consistency of drug convictions are no longer political imperatives, and the costs of maintaining prisons are causing consternation.[25] Voters in Arizona and California passed legislation to reduce sentencing for low-level drug offenders.[26] This may seem like sound policy, given the tenuous relationship between sentencing time and deterrence, but a new study produced by RAND shows that this policy move might be ill advised, too.[27] During the last twenty years, prosecutors have altered their behavior to adapt to the minimum sentencing laws by using them as bargaining power to secure plea bargains.[28] As a result, offenders serving prison time today for low-level drug offenses usually have much more serious criminal histories than their records suggest.[29] Both of the RAND studies have made important contributions to the complex debate on crime and drug policy, and both were made possible by the data commons.[30]

If data anonymity is presumed not to exist, the future of public-use datasets and all of the social utility flowing from them will be thrown into question. Nearly every recent public policy debate has benefited from mass dissemination of anonymized data. Public use data released by the Federal Financial Institutions Examination Council provides a means of detecting housing discrimination and informs

22. JONATHAN P. CAULKINS ET AL., RAND, MANDATORY MINIMUM DRUG SENTENCES: THROWING AWAY THE KEY OR THE TAXPAYERS' MONEY? 62 (1997), *available at* http://www.rand.org/pubs/monograph_reports/MR827.html.

23. *Id.*

24. U.S. SENTENCING COMM'N, SPECIAL REPORT TO THE CONGRESS: MANDATORY MINIMUM PENALTIES IN THE FEDERAL CRIMINAL JUSTICE SYSTEM iii (1991) ("Deterrence, a primary goal of the Sentencing Reform Act and the Comprehensive Crime Control Act, is dependent on certainty and appropriate severity.").

25. K. JACK RILEY ET AL., RAND, JUST CAUSE OR JUST BECAUSE?: PROSECUTION AND PLEA-BARGAINING RESULTING IN PRISON SENTENCES ON LOW-LEVEL DRUG CHARGES IN CALIFORNIA AND ARIZONA xiii (2005), *available at* http://www.rand.org/pubs/monographs/MG288.html.

26. Substance Abuse and Crime Prevention Act of 2000, Cal. Prop. 36 (codified at CAL. PENAL CODE § 1210 (West 2006)); Act Relating to Laws on Controlled Substances and those Convicted of Personal Use or Possession of Controlled Substances, Prop. 200, (Ariz. 1996) (codified as amended at ARIZ. REV. STAT. ANN. § 41–1404.16 (2011).

27. RILEY ET AL., *supra* note 25, at 76.

28. *Id.* at 62.

29. *Id.* at 76.

30. The 1997 study used data from the U.S. Drug Enforcement Agency's System to Retrieve Information from Drug Evidence ("STRIDE") and from the National Household Survey on Drug Abuse. CAULKINS, *supra* note 22, at 85. The 2005 study used data from the California and Arizona Departments of Corrections. RILEY ET AL., *supra* note 25, at 20, 24.

policy debates over the home mortgage crisis.[31] Research performed by health economists and epidemiologists using Medicare and Medicaid data is now central to the debates about health care reform.[32] Census microdata has been used to detect racial segregation trends in housing.[33] Public-use birth data has led to great advances in our understanding of the effects of smoking on fetuses.[34] Public crime data has been used to reveal the inequitable allocation of police resources based on the socio-economic status of neighborhoods.[35] And the data commons is repeatedly used to expose fraud and discrimination that would not be discoverable or provable based on the experience of a single person.[36]

None of this data would be available to the broad research community under a conception of privacy that abandons hope in anonymization. These datasets are critical to what George T. Duncan calls "Information Justice," which is the fairness that accessible information offers to the general public in the form of knowledge, and offers to individuals in the form of a discoverable and verifiable grievance.[37]

C. Ex Ante Valuation Problems

The value of a research database is very difficult to discern in the abstract, before researchers have had a chance to analyze it. The uncertain value makes it difficult to know when privacy interests ought to succumb to the public interest in data sharing. Paul Schwartz demonstrates the problem when he argues that some types of information do not implicate data privacy: "[S]ome kinds of aggregate in-

---

31. Press Release, Federal Financial Institutions Examination Council (Sept. 8, 2006), *available at* http://www.ffiec.gov/hmcrpr/hm090806.htm; Janneke Ratcliffe & Kevin Park, Written Comments and Supplement to Oral Testimony Provided by Janneke Ratcliffe at the Hearing on Community Reinvestment Act Regulations (Aug. 31, 2010), *available at* http://www.ccc.unc.edu/documents/CRA_written_8.6.2010.v2.pdf.

32. *See, e.g.*, Jacob S. Hacker, Inst. for America's Future, *Public Plan Choice in Congressional Health Plans*, CAMPAIGN FOR AMERICA'S FUTURE (Aug. 20, 2009), http://www.ourfuture.org/files/Hacker_Public_Plan_August_2009.pdf.

33. Casey J. Dawkins, *Recent Evidence on the Continuing Causes of Black-White Residential Segregation*, 26 J. URB. AFF. 379, 379 (2004).

34. Allen J. Wilcox, *Birth Weight and Perinatal Mortality: The Effect of Maternal Smoking*, 137 AM. J. EPIDEMIOLOGY 1098, 1098 (1993).

35. Cate, *supra* note 9, at 14.

36. For example, the data routinely collected by the Equal Employment Opportunity Commission is used to check for statistically significant disparities between racial and gender groups. *See, e.g.*, Paul Meier, Jerome Sacks & Sandy L. Zabell, *What Happened in* Hazelwood*: Statistics, Employment Discrimination, and the 80% Rule*, 1984 AM. B. FOUND. RES. J. 139.

37. George T. Duncan, *Exploring the Tension Between Privacy and the Social Benefits of Governmental Databases*, *in* A LITTLE KNOWLEDGE: PRIVACY, SECURITY AND PUBLIC INFORMATION AFTER SEPTEMBER 71, 82 (2004) (Peter M. Shane, John Podesta & Richard C. Leone eds., Century Foundation 2004).

formation involve pools that are large enough to be viewed, at the end of the day, as purely statistical and thus, as raising scant privacy risks as a functional matter."[38] He cites flu trends as an illustration of this sort of aggregate non-problematic data.[39] But Google's Flu Trends — the fastest and most geographically accurate way to monitor national flu symptoms[40] — only works by collecting *all* Google search queries by IP address.[41] This practice runs afoul of Schwartz's admonition against collecting information without a specific and limited purpose.[42]

Google Flu Trends exemplifies why it is not possible to come to an objective, prospective agreement on when data collection is sufficiently in the public's interest and when it is not.[43] Flu Trends is an innovative use of data that was not originally intended to serve an epidemiological purpose. The program uses data that, in other contexts, privacy advocates believe violates Fair Information Practices.[44] This illustrates a concept understood by social scientists that is frequently discounted by the legal academy and policy-makers: some of the most useful, illuminating data was originally collected for a completely unrelated purpose. Policymakers will not be able to determine in advance which data resources will support the best research and make the greatest contributions to society. To assess the value of research data, we cannot cherry-pick between "good" and "bad" data collection.[45]

Take another example, recently reproduced in the Freakonomics blog. The online dating website OkCupid analyzes all of the information entered by its members to reveal interesting truths about the

---

38. SCHWARTZ, *supra* note 12, at 8.

39. *Id.* at 8, 15.

40. Miguel Helft, *Aches, a Sneeze, a Google Search*, N.Y. TIMES, Nov. 12, 2008, at A1.

41. Miguel Helft, *Is There a Privacy Risk in Google Flu Trends?*, N.Y. TIMES BITS (Nov. 13, 2008, 8:20 PM), http://bits.blogs.nytimes.com/2008/11/13/does-google-flu-trends-raises-new-privacy-risks.

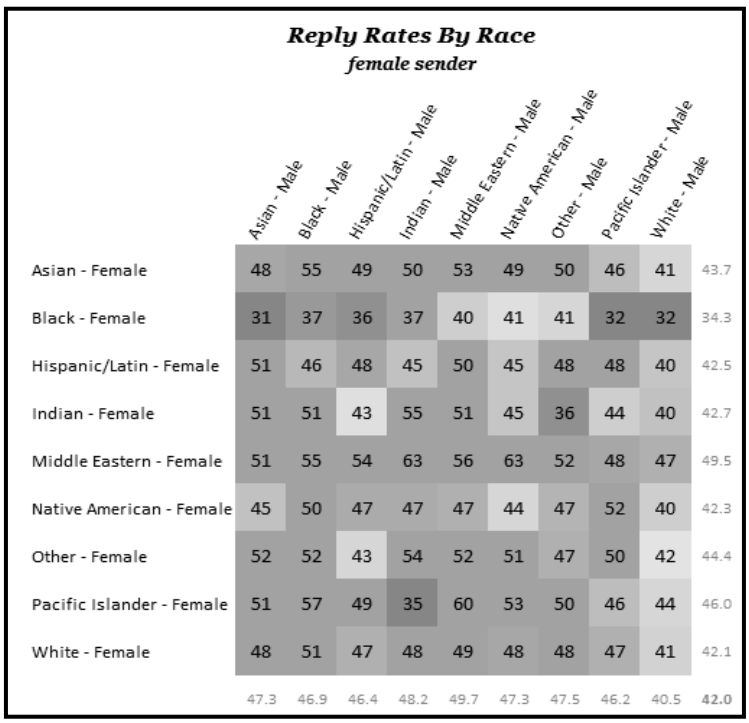
42. SCHWARTZ, *supra* note 12, at 24.

43. The problem of valuing information is as old as privacy. Samuel Warren and Louis Brandeis believed that the press in their day was overstepping "the obvious bounds of propriety and of decency" by photographing the private lives of public and elite figures for the gossip pages. Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 196 (1890). But today gossip journalism is imbedded into mainstream culture and often the spearhead for the uncovering of important news items. *See* David Perel, *How the Enquirer Exposed the John Edwards Affair*, WALL ST. J., Jan. 23, 2010, at A15.

44. *Google Watches as You Type in Search Words and Displays "Live" Results in Real Rime. Creeped Out, So Are We*., TECHALOUD (Aug. 23, 2010), http://www.techaloud.com/2010/08/google-tests-search-results-that-update-as-you-type (expressing displeasure with Google's use of private information in generating search terms); Chris Jay Hoofnagle, *Beyond Google and Evil: How Policy Makers, Journalists and Consumers Should Talk Differently About Google and Privacy*, FIRST MONDAY (Apr. 6, 2009), http://www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2326/2156.

45. *But see* Roger Clarke, *Computer Matching by Government Agencies: The Failure of Cost/Benefit Analysis as a Control Mechanism*, 4 INFO. INFRASTRUCTURE & POL'Y 29 (1995).

dating public.[46] In one fascinating study, the OkCupid researchers found that men of all races responded to the initial contacts of black females at significantly lower rates, despite the fact that the profiles of black females are as compatible as the females of every other race.[47]



Reply Rates By Race

female sender

	Asian - Male	Black - Male	Hispanic/Latin - Male	Indian - Male	Middle Eastern - Male	Native American - Male	Other - Male	Pacific Islander - Male	White - Male	
Asian - Female	48	55	49	50	53	49	50	46	41	43.7
Black - Female	31	37	36	37	40	41	41	32	32	34.3
Hispanic/Latin - Female	51	46	48	45	50	45	48	48	40	42.5
Indian - Female	51	51	43	55	51	45	36	44	40	42.7
Middle Eastern - Female	51	55	54	63	56	63	52	48	47	49.5
Native American - Female	45	50	47	47	47	44	47	52	40	42.3
Other - Female	52	52	43	54	52	51	47	50	42	44.4
Pacific Islander - Female	51	57	49	35	60	53	50	46	44	46.0
White - Female	48	51	47	48	49	48	48	47	41	42.1
	47.3	46.9	46.4	48.2	49.7	47.3	47.5	46.2	40.5	**42.0**

Figure 1: OkCupid Analysis of Member Messaging Behavior[48]

One of the most remarkable aspects of the OkCupid study is that it did not draw the ire of privacy advocates.[49] Contrast Freakonomics's coverage of the OkCupid study with the *L.A. Times*'s coverage of a Facebook study that came to the unsurprising conclusion that Facebook statuses are cheery on holidays and dreary when celebrities die: "If you put something on Facebook, no matter how tight your privacy settings are, Facebook Inc. can still hang onto it, analyze it,

46. *See* Ian Ayres, *Race and Romance: An Uneven Playing Field for Black Women*, FREAKONOMICS, (Mar. 3, 2010, 2:00 PM), http://www.freakonomics.com/2010/03/03/race-and-romance-an-uneven-playing-field-for-black-women.

47. *Id.*

48. *Id.*

49. Its own privacy assurances seemed to have deflected criticism well enough. *See* Jason Del Rey, *In Love with Numbers: Getting the Most out of Your Company Data*, INC. MAGAZINE, Oct. 2010, at 105, 106.

remix it and repackage it. Despite its silly name, the Gross National Happiness indicator is creepy. *We're* in there."[50]

How is it that Facebook's study attracted criticism of its privacy policies while the data used in the OkCupid study went unnoticed? The difference is likely explained by the value of the OkCupid study. The OkCupid study's contribution to our understanding of human relations distracts commentators from thinking about the source of the data. The utility of the research overshadows our collective anxiety about research data. The trouble is that the public and the press undervalue the beneficial uses of research data when the attention turns to data privacy.

The OkCupid study illustrates another important quality of research microdata: that collectively, our data reveals more than any of us could know on our own. The message-writing decisions of each individual OkCupid member could not have revealed the patterns of preferences, but when aggregated, the data supports a hypothesis about human nature and implicit bias. Research data describes everybody without describing anybody. If the data from the OkCupid profiles was thought to be the property of the members, subject to their exclusive determination on the uses to which it is put, society at large, and OkCupid members in particular, would be deprived of the discovery of this quiet pattern.

D. The Importance of Broad Accessibility

The value of data is not completely lost on privacy law scholars, but the need for broad access generally is. When data can be shared freely, it creates a research dialog that cannot be imitated through restricted data and license agreements. In contrast to legal scholars, technology journalists recognize the unmatchable virtues that come from crowdsourcing when all interested people have unfettered access to data.[51] General access ensures the best chance that a novel or creative use of a dataset will not be missed.

Privacy laws that constrain the dissemination of the most useful data through discretionary licensing agreements (such as HIPAA and FERPA) are designed without sufficient appreciation as to how research works. Ironically, they operate on a model that gives researchers too much credit, and has too much faith that data supports just one unassailable version of the truth. In practice, transparency and data sharing are integral to a researcher's credibility. The data commons

50. Mark Milian, *Facebook Digs Through User Data and Graphs U.S. Happiness*, L.A. TIMES TECH. (Oct. 6, 2009, 3:50 PM), http://latimesblogs.latimes.com/technology/2009/10/facebook-happiness.html.

51. *See, e.g.*, *Of Governments and Geeks*, ECONOMIST, Feb. 6, 2010, at 65; Chris Soghoian, *AOL, Netflix and the End of Open Access to Research Data*, CNET SURVEILLANCE STATE (Nov. 30, 2007, 8:30 AM), http://news.cnet.com/8301-13739_3-9826608-46.html.

protects the public discourse from two common research hazards: (1) the failure to catch innocent mistakes, which are legion, and (2) the restriction of access to highly useful data based on ideological considerations or self-interest.

Replication is indispensable to the process of achieving credible, long-lasting results.[52] Just as mistakes and even fabrications occur in the hard sciences,[53] they also occur in the social sciences. The gatekeepers at peer-reviewed science and economics journals have proven to be significantly less effective than the motivated monitoring of peers and foes in the field.[54] For example, a study published in England's preeminent health research journal claimed to have found statistical proof that women can increase the chance of conceiving a male fetus if they eat breakfast cereal.[55] The findings were covered by the *New York Times* and National Public Radio.[56] When the data was made available to other researchers, the results quickly fell apart and have become something of a cautionary tale against researchers that torture a dataset into producing statistically significant results.[57] Simple coding errors are even more common and can distort and completely invert results. Because of the frequency and inevitability of these sorts of errors, the most respected journals make data sharing a prerequisite for publication (and even article submission).[58]

Consider the debate on the deterrent effects of the death penalty. In 1972, the U.S. Supreme Court determined that existing death penalty statutes and practices violated convicts' Eighth Amendment right to

---

52. *See* Gary King, *Replication, Replication*, 28 PS: POL. SCI. & POLITICS 444, 444 (1995).

53. *See Spectacular Fraud Shakes Stem Cell Field*, MSNBC (Dec. 23, 2005), http://www.msnbc.msn.com/id/10589085/ns/technology_and_science-science.

54. The National Institute of Health found that only one out of every twenty claims flowing from observational studies ends up being reproducible in controlled studies. S. Stanley Young, *Everything Is Dangerous: A Controversy*, AM. SCIENTIST (Apr. 22, 2009), http://www.americanscientist.org/science/pub/everything-is-dangerous-a-controversy.

55. Fiona Mathews, et al., *You Are What Your Mother Eats: Evidence for Maternal Preconception Diet Influencing Foetal Sex in Humans*, 275 PROC. ROYAL SOC'Y B 1661, 1665 (2008).

56. Tara Parker-Pope, *Boy or Girl? The Answer May Depend on Mom's Eating Habits*, N.Y. TIMES WELL (April 23, 2008, 12:59 PM), http://well.blogs.nytimes.com/2008/04/23/boy-or-girl-the-answer-may-depend-on-moms-eating-habits; Allison Aubrey, *Can a Pregnant Woman's Diet Affect Baby's Sex?*, (NPR radio broadcast Jan. 15, 2009), *available at* http://www.npr.org/templates/story/story.php?storyId=99346281.

57. *See* Young, *supra* note 54.

58. NATIONAL RESEARCH COUNCIL OF THE NATIONAL ACADEMIES, SHARING PUBLICATION-RELATED DATA AND MATERIALS: RESPONSIBILITIES OF AUTHORSHIP IN THE LIFE SCIENCES 3 (2003). *Science*, an academic journal, changed its review policy in 2006 to require all authors to post the raw data supporting their findings online after the discovery that one of the most important stem cell research findings at that time was a complete fabrication. *See* Barry R. Masters, Book Review, 12 J. BIOMEDICAL OPTICS 039901-1, 039901-1 (2007) (reviewing ADIL E. SHAMOO & DAVID B. RESNIK, RESPONSIBLE CONDUCT OF RESEARCH (2003)).

be free from cruel and unusual punishment.[59] But three years later, an explosive empirical study by Isaac Ehrlich concluded that each execution had the effect of saving up to eight lives by deterring would-be criminals from killing.[60] Robert Bork, then the Solicitor General, cited to Ehrlich's study in his brief for *Gregg v. Georgia*[61] a year later and, lo and behold, the Supreme Court was persuaded to end the moratorium on death sentences.[62] The trouble is, Ehrlich's persuasive study has not stood the test of time and replication. Since then, the capital punishment debate has attracted the attention of many prized economists.[63] John J. Donohue and Justin Wolfers have shown that the empirical studies finding a deterrent effect are highly sensitive to the choice of sampling periods and other discretionary decisions made by the studies' authors.[64] The deterrent effects found by Ehrlich are in doubt, now that economists have had the opportunity to test the robustness of the findings and explore the idiosyncratic series of methodological decisions that led to them.[65] Had Ehrlich alone had access to the crime data supporting his research, and had his study been left to circulate in the media unchallenged, we might not have seen the wane in public and political support for capital punishment that we do today.[66]

Data, just like any other valuable resource, can and often does fall into the control of people or organizations that are politically entrenched.[67] Because the legitimacy of discretionary access decisions is not independently scrutinized, restricted access policies allow data producers to withhold information for politically or financially moti-

---

59. Furman v. Georgia, 408 U.S. 238, 240 (1972).

60. Isaac Ehrlich, *The Deterrent Effect of Capital Punishment: A Question of Life and Death*, 65 AM. ECON. REV. 397, 398 (1975).

61. 428 U.S. 153 (1976).

62. *Id.* at 233–34.

63. *See* John J. Donohue & Justin Wolfers, *Uses and Abuses of Empirical Evidence in the Death Penalty Debate*, 58 STAN. L. REV. 791, 793 (2005) (noting that Lawrence Katz, Steven Levitt, Ellen Shustorovich, Hashem Dezhbakhsh, Paul H. Rubin, Joanna M. Shepherd, H. Naci Mocan, R. Kaj Gittings, and Paul R. Zimmerman have written on the issue).

64. *Id.* at 794. Moreover, with so few capital sentences per year the deterrence effects of each capital sentence cannot be disentangled from the year and state controls. *Id.*

65. The Donohue and Wolfers study has been praised by independent reviewers for its use of sensitivity analysis, and for testing findings against alternative specifications and controls. Joshua D. Angrist & Jörn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics* 15 (Nat'l Bureau of Econ. Research, Working Paper No. 15794, 2010), *available at* http://ssrn.com/abstract=1565896.

66. Steve Chapman, *The Decline of the Death Penalty*, CHI. TRIB., Dec. 26, 2010, at C29; Andrew Kohut, *The Declining Support for Executions*, N.Y. TIMES, May 10, 2001, at A33. The empirical research community has seen a similar debate play out in the context of the gun control debate. *See* Ian Ayres & John J. Donohue III, *Shooting Down the "More Guns, Less Crime" Hypothesis*, 55 STAN. L. REV. 1193, 1202 (2003).

67. This phenomenon is, in fact, what motivates George T. Duncan's concept of "information injustice." Duncan, *supra* note 37, at 71, 82.

vated reasons.[68] A thriving public data commons serves the primary purpose of facilitating research, but it also serves a secondary purpose of setting a data-sharing norm so that politically motivated access restrictions will stick out and appear suspect. Thus, if an entity shared data with researchers under a restricted license to support a study that yielded results that happened to harmonize with the entity's self-interest (as was the case when a pharmaceutical company withheld the raw data from its clinical trials even though the results were used to support an application for FDA approval[69]), the lack of transparency would be a signal that the research may have been tainted by significant pressure to come out a particular way.

Today we get the worst of both worlds. Data can be shared through licensing agreements to whomever the data producer chooses, and privacy provides the agency with an excuse beyond reproach when the data producer prefers secrecy to transparency. This is precisely what happened in *Fish v. Dallas Independent School District*.[70] The Dallas School District denied a request from the Dallas chapter of the NAACP for longitudinal data on Iowa Test scores that would have tracked Dallas schoolchildren over an eleven-year period.[71] Based on expert testimony that a malfeasor could "trace a student's identification with the information requested by [the NAACP] using a school directory," the requested data was found to violate FERPA.[72]

The *Fish* opinion interprets and enforces the FERPA regulations properly.[73] The outcome is consistent with FERPA's statutory goals.

---

68. *See* Lawrence O. Gostin, *Health Services Research: Public Benefits, Personal Privacy, and Proprietary Interests*, 129 ANNALS OF INTERNAL MED. 833 (1998).

69. Pub. Citizen Health Research Grp. v. FDA, No. Civ.A. 99-0177(JR), 2000 WL 34262802, at *1 (D.D.C. Jan. 19, 2000) (G.D. Searle & Co. intervened to support the government's decision to withhold clinical trial data based on the privacy exemption in the FOIA statute).

70. 170 S.W.3d 226 (Tex. App. 2005).

71. *Id.* at 227.

72. *Id.* at 230.

73. The requested dataset would have included the sex, age, ethnicity, random teacher code, random school code, test scores, and a few other variables for each student. The request would have revealed PII because the random school and teacher codes, though they sound like *non-identifiers*, are actually *indirect identifiers*. First, the school codes in the Dallas dataset could be cracked using publicly available school enrollment statistics. For example, if Preston Hollow Elementary School was the only school that enrolled 750 students in the year 1995, then its school code could easily be identified by finding the school in the dataset with 750 subjects for the year 1995. Even if two schools happened to have identical enrollment figures for one particular year, the enrollment patterns over time were unique for every school. (The plaintiffs asked for several consecutive years of test scores.) Once the school codes were reverse-engineered, most of the teacher codes could be re-identified using the same methods. Once the school and teacher codes were cracked, Dallas schoolchildren could be organized into small class clusters. A class of thirty schoolchildren cannot be diced into racial groups and gender categories without dissolving into unique cases. *Cf. infra* Part III. This protocol, checking to see whether subgroups of individuals in a dataset could be re-identified using combinations of publicly documented characteristics, is consistent with the directives promulgated by the Family Policy Compliance Office ("FPCO"), the federal agency charged with enforcing FERPA. In providing guidance on the

However, it also exposes the troubling, draconian results of modern data privacy policy. The data requested by the NAACP might have exposed evidence of discrimination or disparate resource allocation. The school district had the option to cooperate with the NAACP's request by using FERPA's research exemption and providing the data under a restrictive license.[74] Alternatively, the district could have provided a randomized sample of the data so that class sizes could not be used to trace identities. But they had little incentive to do either, and perhaps even an incentive *not* to do so. Privacy law provided the school district with a shield from public scrutiny, and allowed the school district to flout the objectives of public records laws.

We will never know what the *Fish* data might have revealed. Perhaps theories of disparate treatment across class or race lines would have been borne out. Perhaps the research would have facilitated some other, unanticipated finding. Even the confirmation of a null hypothesis can have significant implications, particularly where a portion of the population suspects it may be receiving inequitable treatment. Since privacy law allowed the data producer to avoid disclosure, the value of the withheld data will be forever obscured, and any systemic patterns will be known only to the Dallas school district — if they are known at all. The *Fish* case nicely illustrates the dangers of assigning too little value to research data in the abstract.

E. Freedom of Information Act Requests: Privacy as an Evasion Technique

We would expect public agencies, which are subject to strong public access obligations from FOIA and state public records statutes,[75] to have fewer opportunities to make improperly motivated access decisions. After all, one of the primary goals of public access statutes is to take decisions about who does and does not get to access information out of the hands of the agency.[76] But increased anxieties over the theoretical risk of re-identification arm government agencies with a pretext for denying records requests. As Douglas Sylvester and

---

scope of "personally identifiable information," the FPCO opined that under certain circumstances "the aggregation of anonymous or de-identified data into various categories could render personal identity 'easily traceable.' In those cases, FERPA prohibits disclosure of the information without consent." *See* Letter from LeRoy S. Rooker, Director, Family Policy Compliance Office, to Corlis P. Cummings, Senior Vice Chancellor for Support Services, Bd. of Regents of the Univ. Sys. of Ga. (Sept. 25, 2003), *available at* http://www2.ed.gov/policy/gen/guid/fpco/ferpa/library/georgialtr.html.

74. 20 U.S.C. § 1232g(b)(1)(F) (2006).

75. *See, e.g.*, Freedom of Information Act, 5 U.S.C. § 552 (2006); California Public Records Act, CAL. GOV'T CODE §§ 6250 et seq. (West 2008); Freedom of Information Law, N.Y. PUB. OFFICERS LAW §§ 84 et seq. (Consol. 2011).

76. *See, e.g.*, CAL. GOV'T CODE § 6250 (West 2011) ("[A]ccess to information concerning the conduct of the people's business is a fundamental and necessary right of every person in this state.").

Sharon Lohr have noted, "the strengthening of individual rights-based privacy has allowed some agencies to use privacy as a 'shield' to prevent otherwise appropriate disclosures."[77] The moral hazard reached its apex under the Bush Administration, which shielded the records of current and past presidents from FOIA requests through executive order.[78] The exemption was voluntarily repealed in 2009.[79]

This is not to say that every denial of a public records request is made in bad faith. A number of structural problems plague the process and encumber disclosure. First, the lack of comprehensible standards for privacy protocols (discussed at length in Part V) will tend to drive state agencies to withhold data from researchers if disclosure exposes the agency to liability or sanction. Moreover, the penalties and public criticism for releasing ineffectively anonymized information are much harsher than the consequences of improperly denying a public records request.[80] The imbalanced structural incentives obscure and exacerbate the potential for self-serving behavior. Freedom of information advocates and professional journalism associations allege that privacy exemptions, like national security exemptions, are abused when the requested information is embarrassing for the agency.[81] Thus, as the Society of Professional Journalists puts it, rich data is disclosed about tomato farming and transportation, while data that could be used to vet a government program or expose agency wrongdoing is redacted into oblivion — if it is released at all.[82]

Numerous examples from the FOIA case law support these observations. The Department of Agriculture used the privacy exemption of FOIA to deny a request for the identity of a corporation that compensated or bribed a member of the Dietary Guidelines Advisory Committee.[83] The State Department refused to release documents about forcibly repatriated Haitian refugees to human rights groups — purportedly to protect their privacy.[84] Privacy was "feebly" held up as a justification for declining to collect information about the religious exercise of Navy personnel, in an attempt to rebut a group of Navy chaplains' allegations that nonliturgical Christians were disfavored and underrepresented in the Navy's decisions about hiring, promotion,

77. Sylvester & Lohr, *supra* note 12, at 190; *see also* Cate, *supra* note 9, at 13–15.

78. Further Implementation of the Presidential Records Act, Exec. Order No. 13233, 66 Fed. Reg. 56,025 (Nov. 5, 2001).

79. Presidential Records, Exec. Order No. 13489, 74 Fed. Reg. 4669 (Jan. 21, 2009).

80. For example, in Arizona, improper disclosure of private facts is a felony, while improper denial of a legitimate public records request is a misdemeanor. *See Air Talk: The "Open Government Plan"* (Southern California Public Radio broadcast Dec. 14, 2009), *available at* http://www.scpr.org/programs/airtalk/2009/12/14/the-open-government-plan.

81. *Id.*

82. *Id.*

83. Physicians Comm. for Responsible Med. v. Glickman, 117 F. Supp. 2d 1, 5–6 (D.D.C. 2000).

84. U.S. Dep't of State v. Ray, 502 U.S. 164, 166 (1991).

and retention.[85] In each of these examples, the government's privacy argument eventually failed. But sometimes this argument prevails.[86] And a great majority of denials of public records requests are not litigated at all.[87]

In 2008, UCLA denied a public records request that a faculty member on the undergraduate admissions committee submitted for the University's admissions data.[88] UCLA concluded that the request posed "serious privacy concerns" and could not be fulfilled without violating FERPA.[89] Astonishingly, the same rationale did not impede UCLA from sharing similar admissions data under a restricted license agreement to a different UCLA professor.[90] The only appreciable difference between the two requests was the divergent attitudes each professor maintained toward UCLA's admissions process. The denied requester openly questioned whether the school was using applicant race information in an impermissible way.[91]

The University of Arkansas Little Rock ("UALR") School of Law denied a similar request for admissions data from a faculty member on its admissions committee. The professor regularly reviewed the original, raw admissions files, but the school denied his request for data, claiming that FERPA prohibited the release of even de-identified statistical data.[92] When a UALR Law School alumna requested access to similar application data in an independent request, the University (perhaps inadvertently) disclosed a memorandum of notes documenting advice from their legal counsel: "We say FERPA, they can challenge if they want."[93] A cogent interpretation is that the federal privacy law is being used as a tactical device to greatly increase the transaction costs for public records requests. Since requests for anonymized university and law school admissions data have already passed judicial scrutiny assessing FERPA compliance,[94] the general

85. Adair v. England, 183 F. Supp. 2d 31, 56 (D.D.C. 2002).

86. *See* Fish v. Dallas Indep. Sch. Dist., 170 S.W.3d 226 (Tex. App. 2005).

87. COALITION OF JOURNALISTS FOR OPEN GOV'T, FOIA LITIGATION DECISIONS, 1999–2004 1 (2004), *available at* http://www.cjog.net/documents/Litigation_Report_9904.pdf.

88. TIMOTHY GROSECLOSE, CUARS RESIGNATION REPORT (2008), *available at* http://images.ocregister.com/newsimages/news/2008/08/CUARSGroseceloseResignationReport.pdf; *see also* Seema Mehta, *UCLA Accused of Illegal Admitting Practices*, L.A. TIMES, Aug. 30, 2008, at B1.

89. *See* GROSECLOSE, *supra* note 88.

90. *Id.*

91. *Id.*

92. Robert Steinbuch, What They Don't Want Me (and You) to Know About Non-Merit Preferences in Law School Admissions: An Analysis of Failing Students, Affirmative Action, and Legitimate Educational Interests 3 (unpublished manuscript) (on file with author).

93. Richard J. Peltz, *From the Ivory Tower to the Glass House: Access to "De-Identified" Public University Admission Records to Study Affirmative Action*, 25 HARV. J. ON RACIAL & ETHNIC JUSTICE 181, 185–87 (2009).

94. *See, e.g.*, Osborn v. Bd. of Regents of Univ. of Wis., 647 N.W.2d 158, 171 (Wis. 2002) ("[B]y redacting or deleting the name of the high school or undergraduate institution, the University no longer faces a situation where only one minority student from a named

counsel's offices at UCLA and UALR ought to have known that, with minimal effort, a sufficiently safe admissions dataset could be produced.

The distribution of access to data is a problem worthy of national attention and concerted effort. The data commons is a powerful, natural antidote to information abuses. It is critical for information justice, since our pooled data can reveal the patterns of human experience that no single anecdote can. Since the value of a dataset cannot be determined ex ante, any rule that significantly impedes the release of research data imposes a social cost of uncertain magnitude.

III. DOOMSDAY DETECTION: THE COMPUTER SCIENCE APPROACH

A large body of computer science literature explores the theoretical risk that a subject in an anonymized dataset can be re-identified. De-anonymization scientists study privacy from an orientation that emphasizes any harm that is theoretically possible. They are in the habit of looking for worst-case scenario risks.[95] This orientation grows out of a natural inclination to believe that, if there is value to abusing anonymized data, and if re-identification is not too difficult, then such re-identification will happen. In other words, where there is motive and opportunity, a de-anonymization attack is a foregone conclusion. The de-anonymization scientists' perspective has some intuitive appeal, and the legal literature has embraced the findings and predictions of the computer science literature without much skepticism.[96] The de-anonymization literature taps into privacy advocates' natural unease any time information is distributed without the consent of the data subjects.

In this Part, I briefly explain how de-anonymization attacks work.[97] Next, I explore the lessons growing out of the computer science literature and find that they greatly exaggerate the opportunities and motivations of the hypothetical adversary. The computer science

high school applies to one of the University's campuses and therefore, even though the student's name is not disclosed, the data could be personally identifiable.").

95. Mark Elliot, *DIS: A New Approach to the Measurement of Statistical Disclosure Risk*, 2 RISK MGMT. 39 (2000) (putting forward a new method of measuring the "worst-case risk"); Jordi Nin et al., *Rethinking Rank Swapping to Decrease Disclosure Risk*, 64 DATA & KNOWLEDGE ENGINEERING 346 (2008). But note that many computer scientists also incorporate assessments of data utility and information loss into their work. *See, e.g.*, DUNCAN, *supra* note 11; Josep Domingo-Ferrer et al., *Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk*, EUROPEAN COMMISSION (2001), http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/81.pdf.

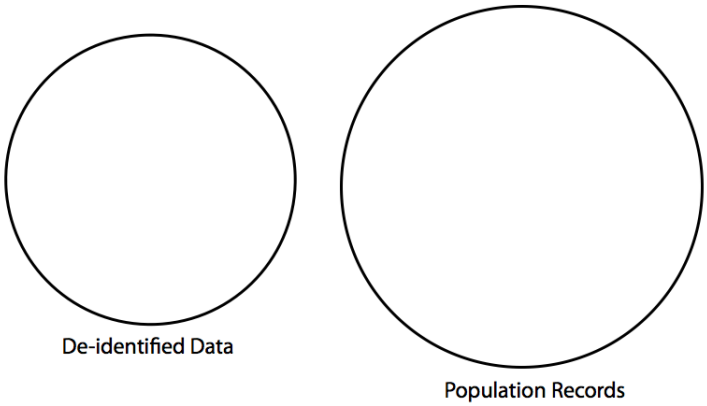
96. *See infra* notes 117–119, 135 and accompanying text.

97. For a concise overview on how de-anonymization attacks work, see JANE YAKOWITZ & DANIEL BARTH-JONES, TECH. POLICY INST., THE ILLUSORY PRIVACY PROBLEM IN *SORRELL V. IMS HEALTH* 1–5 (2011), http://www.techpolicyinstitute.org/files/the%20illusory%20privacy%20problem%20in%20sorrell1.pdf.

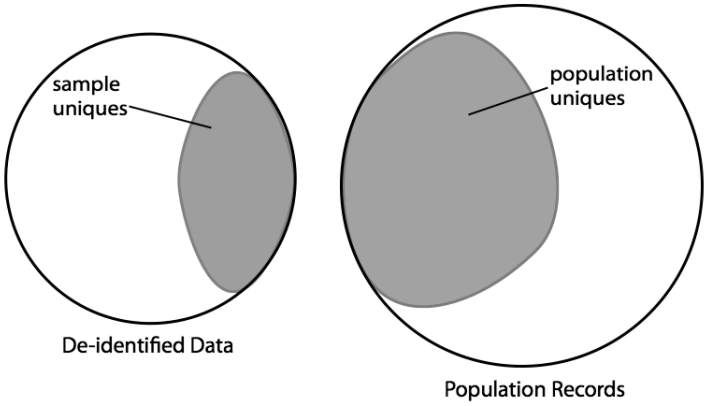
literature (and the policymakers who borrow from it) makes five inaccurate assertions: (1) every variable in a dataset is an indirect identifier; (2) data supporting inferences about a population of data subjects violates privacy; (3) useful data is necessarily privacy-violating; (4) re-identification techniques are easy; and (5) public datasets have value to an adversary over and above the information he already has. I will address each of these in turn.

A. How Attack Algorithms Work

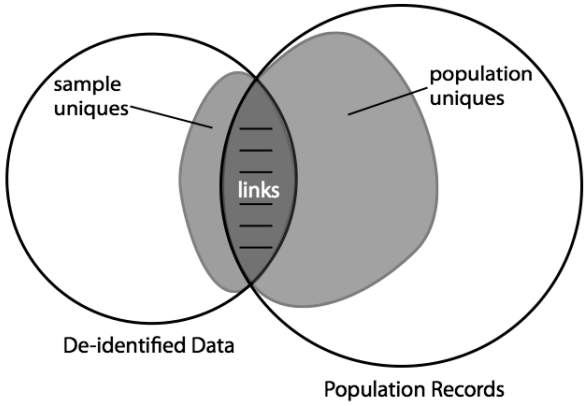
All de-anonymization attack algorithms are variants of one basic model. An adversary attempts to link subjects in a de-identified database to identifiable data on the entire relevant population ("population records"). The adversary links the two databases using indirect identifier variables that the two datasets have in common. To visualize the attack, suppose the two circles in this diagram represent the indirect identifiers in the de-identified database and the population records, respectively. Initially, these databases have no linkages:



The adversary identifies subjects in the de-identified data that have a unique combination of values among the indirect identifiers. He does the same to the population records:



Finally, the adversary links all the sample uniques he can to the population uniques:



Only a subset of the sample uniques and population uniques will be linkable because some of the sample uniques might not actually be unique in the population, and some of the population uniques might not be present in the sample of the de-identified data.[98]

98. More sophisticated techniques will make matches not based on strong exact linkages but on the similarity of the matching variables and the greater deviation between the best match and the second-best match. This allows an attack algorithm to make matches under more realistic conditions in which databases contain measurement error, but it nevertheless requires that the adversary have access to more-or-less complete information on the general

Latanya Sweeney provided the classic example of a successful matching attack when she combined de-identified Massachusetts hospital data with identifiable voter registration records in order to re-identify Governor William Weld's medical records.[99] Because the hospital data at that time — before the passage of HIPAA — included granular detail on the patients (5-digit ZIP code, full birth date, and gender), many patients were unique in the hospital data and the voter records.

Today, there is little disagreement that this sort of "trivial de-identification" of records — the removal of only direct identifiers like names, social security numbers, and addresses — is insufficient on its own. Subjects can too easily be identified through a combination of indirect identifiers. Thus, like other federal privacy statutes, HIPAA requires data producers to remove not only the obvious direct identifiers, but also *any* information known by the disclosing agency that can be used alone or in combination with other information to identify an individual subject.[100]

While there is broad agreement on the rejection of trivial de-identification, privacy experts disagree on the efficacy of current best practices. Legal scholars and advocacy groups limit their focus to the computer science studies falling on one side of the debate — those making the common erroneous assertions explored below — while ignoring the disclosure-risk research coming out of the statistical and public health disciplines. This has had the unfortunate consequence of leading the legal and policy discourse astray.

B. Erroneous Assertions

The mounting literature on privacy risks associated with anonymized research data propagates five myths about re-identification risk. In combination, these inaccurate assertions lead lay audiences to believe that anonymized data cannot be safe.

1. Not Every Piece of Information Can Be an Indirect Identifier

Disclosure risk analysis has traditionally looked for categories of information previously disclosed to the public in order to distinguish "indirect identifiers" from "non-identifiers." For example, data subjects' names and addresses are available in voter registration rosters

---

population from which the de-identified data was sampled. These methods are described more thoroughly by Josep Domingo-Ferrer et al., *supra* note 95, at 813–14.

99. *See* Sweeney, *supra* note 11, at 52.

100. HIPAA Standards for Privacy of Individually Identifiable Health Information, 45 C.F.R. § 164.514(b)(2)(ii) (2010). Alternatively, the disclosing entity must use "generally accepted statistical and scientific principles and methods" to ensure that the risks of re-identification are "very small." § 164.514(b)(1).

(which are public records); therefore ZIP codes and other geographic codes must be classified as indirect identifiers.[101] On the other hand, food preferences are not systematically collected and re-released publicly, so a variable describing the subject's favorite food would traditionally be considered a non-identifier.

De-anonymization scientists do not limit the theoretical adversary to public sources of information. The most influential de-anonymization study, by Arvind Narayanan and Vitaly Shmatikov, describes the re-identification of subjects in the Netflix Prize Dataset.[102] In 2006, Netflix released an anonymized dataset to the public consisting of movie reviews of 500,000 of its members.[103] Narayanan and Shmatikov used information from user ratings on the Internet Movie Database (IMDb) to re-identify subjects in the Netflix Prize dataset.[104] This study is regarded as proof that publicly accessible datasets can be reverse-engineered to expose personal information even when state-of-the-art anonymization techniques are used.[105] The study energized the press because the auxiliary information Narayanan and Shmatikov used was collected from the Internet. But before diving into how the algorithm works, it is helpful to note a chasm between Narayanan and Shmatikov's conception of privacy risk and that enshrined in U.S. privacy statutes.

Narayanan and Shmatikov examine how auxiliary information learned through any means at all, even at the water cooler, could be used to identify a target.[106] They ask, "if the adversary knows a few of the [target] individual's purchases, can he learn *all* of her purchases?" and "if the adversary knows a few movies that the individual watched, can he learn *all* movies she watched?"[107] The implicit directive from these questions is that public datasets must be immune from targeted attacks using special information. The belief that privacy policy is expected to protect data even from snooping friends and coworkers is

---

101. *What is a Quasi-identifier?*, ELECTRONIC HEALTH INFO. LABORATORY (Oct. 18, 2009), http://www.ehealthinformation.ca/knowledgebase/article/AA-00120. Note that indirect identifiers are also known as "quasi-identifiers."

102. Narayanan & Shmatikov, *supra* note 11.

103. *Id.*

104. *Id.* at 122–23. The authors first mapped the five-point scale from Netflix movie ratings onto the ten-point scale used by IMDb, and then attempted to identify matches based on strings of movies that were reviewed similarly on both websites. *Id.*

105. Brief of *Amicus Curiae* Electronic Frontier Foundation in Support of Petitioners at 9–10, Sorrell v. IMS Health Inc., 131 S.Ct. 2653 (2011) (No. 10-779), 2011 WL 757416, at *9–10; *see also supra* note 5 (discussing various privacy lawsuits). Netflix had added random noise to the dataset. Narayanan & Shmatikov, *supra* note 11, at 119.

106. *See* Narayanan & Shmatikov, *supra* note 11, at 122 ("A water-cooler conversation with an office colleague about her cinematographic likes and dislikes may yield enough information [to de-anonymize her subscriber record] . . . .").

107. *Id.* at 112; *see also* Cynthia Dwork, *Differential Privacy*, 2006 PROC. 33RD INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES & PROGRAMMING, *available at* http://research.microsoft.com/pubs/64346/dwork.pdf.

adopted reflexively by Paul Ohm without acknowledging that it introduces a significant departure from the design of current law: "To summarize, the next time your dinner party host asks you to list your six favorite obscure movies, unless you want everybody at the table to know every movie you have ever rated on Netflix, say nothing at all."[108] If public policy had embraced this expansive definition of privacy — that privacy is breached if somebody in the database could be re-identified by anybody else using special non-public information — dissemination of data would never have been possible. Instead, U.S. privacy law in its various forms requires data producers to beware of indirect identifiers that are, or foreseeably could be, in the public domain.[109]

However, Narayanan and Shmatikov's study has sway because the Internet gives a malfeasor access to more information than he ever had before. Narayanan and Shmatikov were able to use the IMDb movie reviews of two strangers to re-identify them in the Netflix data.[110] Their study illustrates how the Internet is a (relatively) new public information resource that blurs the distinction between non-identifiers and indirect identifiers.[111] The Internet affects data anonymization by archiving and aggregating large quantities of information and by making information gathering practically costless.[112] It also provides a platform for self-revelation and self-publication, making the available range of information about any one person unpredictable and practically limitless.

Current privacy policy does not anticipate how we should deal with this shift. On one hand, if anybody can access information on the Internet, it seems unquestionable that the information is "public." Thus, this information might best be described as an indirect identifi-

---

108. Ohm, *supra* note 4, at 1721.

109. For example, regulations issued under FERPA define PII to include "information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, *who does not have personal knowledge of the relevant circumstances*, to identify the student with reasonable certainty." 34 C.F.R. § 99.3 (2011) (emphasis added). Likewise, "[a]t a minimum, each statistical agency must assure that the risk of disclosure from the released data when combined with other relevant *publicly available* data is very low." *Report on Statistical Disclosure Limitation Methodology* 3 (Fed. Comm. on Statistical Methodology, Statistical Working Paper No. 22, 2d version, 2005) [hereinafter Working Paper No. 22] (emphasis added), *available at* http://www.fcsm.gov/working-papers/SPWP22_rev.pdf.

110. Narayanan & Shmatikov, *supra* note 11, at 123.

111. *See generally id.* Narayanan and Shmatikov make similar breakthroughs using graphs of network connections of anonymized Twitter accounts by matching them to sufficiently unique networked accounts on Flickr. Arvind Narayanan & Vitaly Shmatikov, *Deanonymizing Social Networks*, 2009 PROC. 30TH IEEE SYMP. ON SECURITY & PRIVACY 173.

112. Schwartz, *supra* note 6; Daniel J. Solove, *Access and Aggregation: Public Records, Privacy and the Constitution*, 86 MINN. L. REV. 1137, 1185 (2002) ("The aggregation problem arises from the fact that the digital revolution has enabled information to be easily amassed and combined.").

er. On the other hand, data sharing will be severely constrained if the status of a category of information is shifted from non-identifier to indirect identifier simply because members of a small minority of data subjects choose to reveal information about themselves. If I blog about a hospital visit, should my action render an entire public hospital admissions database (relied on by epidemiologists and health policy advocates) in violation of privacy law? Are the bounds of information flow really to be determined by the behavior of the most extroverted among us?[113] This looks like a quagmire from which no reasonable normative position can emerge.[114] The approach that I endorse in Part V sidesteps this question because the issue does not become relevant until we reach the apocalyptic scenario in which reidentification is a plausible risk, and adversaries painstakingly troll through our blogs to put together complete dossiers. For reasons that will soon become evident, such adversaries are unlikely to materialize.

The Netflix study makes an excellent contribution to our knowledge base, but it is a theoretical contribution. The Narayanan-Shmatikov de-anonymization algorithm is limited to a set of anonymized datasets with particular characteristics. For the algorithm to work, the dataset must be large (in the sense of having a large number of variables or attributes), and it must be sparse (which is a technical term roughly meaning that most of the dataset is empty, and that the data subjects are readily distinguishable from each other).[115] Moreover, because the attack algorithm infers population uniqueness from sample uniqueness, the research dataset must have accurate and complete information about the data subjects in the sample in order to avoid false positives and negatives[116] — a condition that does not

113. As Andrew Serwin puts it, "[i]ndeed, in today's Web 2.0 world, where many people instantly share very private aspects of their lives, one can hardly imagine a privacy concept more foreign than the right to be let alone." Andrew Serwin, *Privacy 3.0 — The Principle of Proportionality*, 42 U. MICH. J. L. REFORM 869, 872 (2009).

114. Indeed, "lifelogging" on the Internet presents a number of challenges for privacy scholars even on their own. Anita Allen has written about the problems of the Internet's "pernicious memory" recalling information that puts the lifelogger in the worst light. Anita L. Allen, *Dredging Up the Past: Lifelogging, Memory, and Surveillance*, 75 U. CHI. L. REV. 47, 56–63 (2008).

115. *See* Narayanan & Shmatikov, *supra* note 11, at 111.

116. The Narayanan-Shmatikov algorithm utilizes the dataset's sparseness to test for false positive matches. If a set of movies leads to a unique match in the Netflix data, and if the movies don't share a common fan base, then the algorithm will be confident that the match is accurate. *Id.* at 112. But the Netflix Data is missing a lot of information about the movie-viewing of its own data subjects. The algorithm is susceptible to false positives and false negatives when it attempts to match against auxiliary information. Take this simplified but illustrative hypothetical: Albert, Bart, and Carl have all seen *Doctor Zhivago*, *Evil Dead II*, and *Dude, Where's My Car?*. Albert and Bart are in the Netflix database, Carl is not. Albert rates all three movies, but Bart rates only *Doctor Zhivago*, and, thus, Netflix has no record of his having seen *Evil Dead II* and *Dude, Where's My Car?*. Because Albert is the only person in the Netflix dataset who rated all three movies, he looks highly unique among

even hold for the Netflix data and is certainly not characteristic of most large commercial datasets, such as consumer data from Amazon. And, importantly, the adversary must understand entropic de-anonymization in order to test the confidence level of his algorithm's match.

These limitations are sizeable, yet they are entirely ignored by the legal scholars, privacy advocates, civil litigants, and now, the FTC, relying on the study to conclude that anonymization is dead.[117] The Narayanan-Shmatikov study has provided the first ping in an echo chamber that has distorted the conversation about public research data. Consider, for example, this report prepared by the preeminent privacy scholar Paul Schwartz:

> Regarding the question of PII versus non-PII, recent work in computer science has shown how easy it can be to trace non-PII to identifiable individuals . . . . [A] study involving Netflix movie rentals was able to identify *eighty percent* of people in a supposedly anonymous database of 500,000 Netflix users; the identification was triggered by their ratings in the Netflix database of at least three films.[118]

The Electronic Privacy Information Center ("EPIC") has gone further, claiming that the study authors re-identified 99 percent of the Netflix users.[119] These statements bear scant relation to reality. In fact, Narayanan and Shmatikov performed a proof of concept study on a small sample of IMDb users. They successfully re-identified two of the IMDb users in the Netflix database.[120] There is a real risk that the

---

the Netflix data subjects, even though we know, in fact, that these three movies are not unique to him even within the Netflix sample. Carl comments on all three movies on IMDb. The attack algorithm matches Carl's IMDb profile to Albert's Netflix data and reports back with a high degree of statistical confidence that the match is not a false positive.

117. In January 2010, a panel of privacy law experts and computer scientists advised the FTC that, in promulgating new regulations, it should abandon faith in anonymization and clamp down on broad data sharing to the extent possible. The Narayanan-Shmatikov study was held up as evidence that anonymization protocols offer no security against re-identification. Remarks at the FTC Second Roundtable on Exploring Privacy 15, 56 (Jan. 28, 2010) (transcript available at http://www.ftc.gov/bcp/workshops/privacyroundtables/PrivacyRoundtable_Jan2010_Transcript.pdf). Narayanan, however, cognizant of the importance of research data, has worked with entities to anonymize public release datasets sufficiently to reduce risks. *See* Steve Lohr, *The Privacy Challenge in Online Prize Contests*, N.Y. TIMES BITS (May 21, 2011, 5:25 PM), http://bits.blogs.nytimes.com/2011/05/21/the-privacy-challenge-in-online-prize-contests.

118. SCHWARTZ, *supra* note 12, at 7 (emphasis added).

119. Brief of Amici Curiae Electronic Privacy Information Center (EPIC) et al. in Support of the Petitioners at 33, Sorrell v. IMS Health, Inc., 131 S. Ct. 2653 (2011) (No. 10-779), *available at* http://www.scotusblog.com/case-files/cases/sorrell-v-ims-health-inc.

120. Narayanan & Shmatikov, *supra* note 11, at 122–23.

echo chamber will continue to distort the reasoned judgment of lawmakers and regulators if such misconceptions are not corrected now.

Of the studies conducted in the last decade, only one was conducted under the conditions that replicate what a real adversary would face while also verifying the re-identifications. The Federal Department of Health and Human Services Office of the National Coordinator for Health Information Technology ("ONC") put together a team of statistical experts to assess whether data properly de-identified under HIPAA can be combined with readily available outside data to re-identify patients.[121] The team began with a set of approximately 15,000 patient records that had been de-identified in accordance with HIPAA.[122] Next, they sought to match the de-identified records with identifiable records in a commercially available data repository and conducted manual searches through external sources (e.g., InfoUSA) to determine whether any of the records in the identified commercial data would align with anyone in the de-identified dataset.[123] The team determined that it was able to accurately re-identify two of the 15,000 individuals, for a match rate of 0.013%.[124] In other words, the risk — even after significant effort — was very small.[125]

Other, less attention-grabbing studies from the field of statistical disclosure risk have similarly differed from the conclusions drawn by the Narayanan-Shmatikov study: in realistic settings, datasets can rarely be matched to one another because both sets of data usually contain substantial amounts of measurement error that decimate the opportunity to link with confidence.[126] This is not the sort of difficulty that can be overcome with technology or shrewd new attack techniques; rather, it is a natural protection afforded by the inherently messy nature of data and of people.[127]

2. Group-Based Inferences Are Not Disclosures

Computer scientists have an expansive definition of privacy. They count as privacy breaches even mere inferences that might be applied to an individual based on subgroup statistics. Justin Brickell and

121. Deborah Lafky, Dep't of Health and Human Servs. Office of the Nat'l Coordinator for Health Info. Tech., The Safe Harbor Method of De-Identification: An Empirical Test 15–19 (2009), http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf.

122. *Id.* at 16.

123. *Id.* at 17–18.

124. *Id.* at 19.

125. These findings are consistent with an earlier study that examined re-identification attacks under realistic conditions. *See* U. Blien et al., *Disclosure Risk for Microdata Stemming from Official Statistics*, 46 STATISTICA NEERLANDICA 69 (1992).

126. *See id.* at 80–81.

127. Even under conditions that are considered risky, re-identification of anonymized datasets is difficult to pull off due to the "natural unreliability of measurement," which serves as a natural barrier. Walter Müller, et al., *Identification Risks of Microdata*, 24 SOC. METHODS & RES. 131, 151 (1995).

Vitaly Shmatikov, computer scientists at the University of Texas whose work has greatly influenced Paul Ohm's scholarship, define privacy breach to include the release of any information where the distribution of a sensitive variable for a subgroup of data subjects differs from that variable's distribution over the entire sample.[128] Similarly, Cynthia Dwork has crafted her definition of "differential privacy" to cover group privacy.[129]

This conception of a privacy right — one that protects against the disclosure of any sensitive information that differs by demographic subgroup — avoids two potential harms that can result from group inference disclosure. First, facts about a group can be used to make a determination about an individual. For example, a health care provider might deny coverage to a member of a particular subgroup based on the health profiles of the entire subgroup. Second, group differences in a sensitive characteristic can lead the public to adopt inappropriate stereotypes that mischaracterize individuals and lead to prejudices. James Nehf describes the problem as so: "Since the information used to form [a] judgment is not the complete set of relevant facts about us, we can be harmed (or helped) by the stereotyping or mischaracterization."[130]

These criticisms are shortsighted. They are, in fact, attacks on the very nature of statistical research. Federal statistical agencies have responded to concerns about subgroup inference disclosure with two persuasive retorts. "First[,] a major purpose of statistical data is to enable users to infer and understand relationships between variables. If statistical agencies equated disclosure with inference, very little data would be released."[131] 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, such as a heightened risk of cancer for a discernable demographic or geographic group, then the data is likely to be useful for exploring and understanding the causes of those differences.[132]

---

128. Justin Brickell & Vitaly Shmatikov, *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*, 2008 PROC. 14TH ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING (KDD) 70, 72; *see also* Narayanan & Shmatikov, *supra* note 11, at 114.

129. Dwork, *supra* note 107, at 9.

130. James P. Nehf, *Recognizing the Societal Value in Information Privacy*, 78 WASH. L. REV. 1, 24 (2003). Similar arguments have arisen in response to the disclosure of information about Tay-Sachs disease in the Jewish community and sickle-cell anemia in the African-American population. Lawrence O. Gostin & Jack Hadley, *Health Services Research: Public Benefits, Personal Privacy, and Proprietary Interests*, 129 ANNALS OF INTERNAL MED. 833, 834 (1998).

131. Working Paper No. 22, *supra* note 109, at 11.

132. I discuss in Part IV how the Brickell and Shmatikov definition of privacy has misled legal scholars to believe that there is a forced choice between privacy and data utility.

"Second, inferences are designed to predict aggregate behavior, not individual attributes, and thus are often poor predictors of individual data values."[133] That is to say, the use of aggregate statistics to judge or make a determination on an individual is often inappropriate. Though stereotyping might happen anyway, it has never been a goal of privacy law to prevent all forms of ignorant speculation. Stereotyping will not go away by suppressing data. To the contrary, data can be very useful in debunking stereotypes.[134]

3. A Data Release Can Be Useful and Safe at the Same Time

Paul Ohm argues that if data is useful to researchers, it must create a serious risk of re-identification.[135] This claim has been repeated in the national media.[136] But the assertion is erroneous. A database with just one indirect-identifying variable (such as gender) tied to non-public information (such as pharmaceutical purchases) can be tremendously valuable for a *specific* research question — such as: "Do women purchase drugs in proportion to the national rates of diagnosis?" — without any risk of re-identification. Ohm and the media outlets were thrown off because the technical studies they cite use a definition of data-mining utility that encompasses *all possible* research questions that could be probed by the original database.[137] So, for example, if race and geographic indicators are removed from the database, the utility of that database for all possible research questions plummets, even though the utility of that database for this specific research question stays intact. For specific research questions, utility and anonymity can and often do coexist.

133. Working Paper No. 22, *supra* note 109, at 11.

134. To the very limited extent group inference privacy has been tested in the courts, judges have been unwilling to recognize an implied contract or privacy challenge to releases of de-identified data, even when the de-identified data could be used to make group inferences for marketing purposes. *See* London v. New Albertson's, Inc., No. 08-CV-1173 H(CAB), 2008 WL 4492642, at *5–6 (S.D. Cal. Sept. 30, 2008) (holding that the disclosure of anonymous individual-level pharmacy patient data to a marketing firm did not contravene assurances from a pharmacy that it "collects your personal information and prescription information only for the fulfillment of your prescription order and to enable you to receive individualized customer service beyond what we can provide to anonymous users").

135. Ohm, *supra* note 4, at 1755.

136. Singel, *supra* note 5.

137. *See* Brickell & Shmatikov, *supra* note 128, at 74. The study does helpfully prove that small increases in privacy protection cause disproportionately large destruction of overall utility. *Id.* at 78. But if privacy protocols are designed to preserve the utility of a dataset for a *particular* research question, nothing in the study suggests that this would not be possible.

4. Re-Identifying Subjects in Anonymized Data Is Not Easy

Computer scientists concerned about data privacy face the challenge of convincing the public that an adversary of low-to-moderate skill is capable of performing the same sort of attacks that they can. De-anonymization scientists often refer to the fact that their attacks can be performed on home computers using popular programs.[138] Paul Ohm makes the same rhetorical move in order to argue that we are living in the era of "easy reidentification."[139]

> The Netflix study reveals that it is startlingly easy to reidentify people in anonymized data. Although the average computer user cannot perform an inner join, most people who have taken a course in database management or worked in IT can probably replicate this research using a fast computer and widely available software like Microsoft Excel or Access.[140]

While the Netflix attack algorithm could be performed using Excel, an adversary would have to understand the theory behind the algorithm in order to know whether the dataset is a good candidate and whether matches should be rejected as potential false positives.[141] The suggestion that anybody with an IT background and a copy of Excel can do this is implausible.

The myth of easy re-identification was tested and rejected in the case of *Southern Illinoisan v. Illinois Department of Public Health*.[142] In that case, the plaintiff newspaper submitted a public records request to the Illinois Department of Public Health for a table containing the ZIP codes, dates of diagnosis, and types of cancer for hospital patients in the department's database.[143] The plaintiff newspaper's goal was to test whether certain forms of cancer were clustered in distinct geographic areas,[144] which would have suggested that their incidence was created or greatly exacerbated by environmental factors.[145] The government relied on the testimony of Dr. Latanya Sweeney to support its argument that granting the request would violate cancer patient privacy because the data could be de-anonymized.[146]

---

138. *See, e.g.*, S. Illinoisan v. Ill. Dep't of Pub. Health, 844 N.E.2d 1, 7 (Ill. 2006).
139. Ohm, *supra* note 4, at 1716.
140. *Id.* at 1730 (footnote omitted).
141. *See supra* text accompanying notes 115–116.
142. 844 N.E.2d 1.
143. *Id.* at 3.
144. *Id.*
145. *See id.* at 7.
146. *Id.* at 4. The privacy standard for this case was heightened from PII to information that "*tends* to lead to the identity." *Id.* at 18 (emphasis added). Nevertheless the court found that the government failed to demonstrate that the requested data would tend to lead to the

Dr. Sweeney's testimony about the process she used to re-identify subjects is under seal out of a fear that the opinion would create an instruction book for a true malfeasor,[147] but the description in the Illinois Supreme Court's opinion suggests that she did the following[148]: She began by researching the disease of neuroblastoma — the rare form of cancer of interest to the plaintiff newspaper — in order to familiarize herself with the symptoms and treatment.[149] Next, she purchased two thousand dollars' worth of public and "semi-public" datasets, some of which required her to fill out forms and wait for processing.[150] Some of these purchased datasets (probably voter registration data) identified their subjects by name and address.[151] If Dr. Sweeney employed the same processes that she had previously used to re-identify health records, it is very likely that she linked the identifiable data to pre-HIPAA hospital discharge data that had not been anonymized (only the names had been removed) by using granular detail about the hospital patients' dates of birth, sex, and ZIP codes.[152] Since the passage of HIPAA, such information is no longer publicly available.[153] Next, Dr. Sweeney used what she learned about neuroblastoma to identify possible neuroblastoma patients in the combined purchased databases.[154] The purchased data contained some information — secondary diagnoses or prescription drug treatments perhaps — that allowed her to infer which people in the consumer databases suffered from neuroblastoma.[155] Since the purchased public data was linked to identities, she was able to use what she learned from the purchased resources to produce accurate names for most of the entries in the requested cancer registry dataset.[156]

Dr. Sweeney testified that it would be very easy for anyone to identify people in the cancer registry dataset:

> It is very easy in the following sense, all I used was commonly available PC technology . . . [a]nd readily available software . . . and all that was required were the simple programs of using [spreadsheets]. . . . They come almost on every machine now days

identities of the subjects. *Id.* at 21. Before she took the witness stand in this case, Dr. Sweeney had demonstrated that re-identification of allegedly anonymized data was possible by reverse-engineering Massachusetts medical data. *See* Sweeney, *supra* note 11.

147. *S. Illinoisan*, 844 N.E.2d at 7–8.
148. *Id.* at 8.
149. *Id.*
150. *Id.*
151. *Id.* at 4.
152. *See* Sweeney, *supra* note 11.
153. 45 C.F.R. § 164.514(b)(2)(i) (2010).
154. *S. Illinoisan*, 844 N.E.2d at 8.
155. *Id.*
156. *Id.*

> [sic] . . . so they don't require you have [sic] any programming or require you to take a computer class, but they do require you to know the basics of how to use the machine and how to use those simple packages.[157]

The Illinois Supreme Court was not convinced. The court reasoned that it was Dr. Sweeney's "'knowledge, education and experience in this area' that made it possible for her to identify the Registry patients" and not merely her access to Microsoft Excel.[158] Because Dr. Sweeney used her well-honed discretion to make matches between two data sources that did not map easily onto each other, Dr. Sweeney's methods took advantage of her efforts and talents. *Southern Illinoisan* and the Netflix example illustrate that designing 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.

5. De-Anonymized Public Data Is Not Valuable to Adversaries

The plaintiffs in *Southern Illinoisan* had a second objection to Dr. Sweeney's testimony: Dr. Sweeney identified neuroblastoma patients using the purchased data resources, not the dataset requested by the plaintiffs.[159] She used the requested table "only to verify her work"[160]: she checked to see if the ZIP codes and diagnosis dates of her neuroblastoma candidate guesses matched the anonymous cancer registry.[161]

The requested table undoubtedly provided some value by allowing her to have more confidence in the attack algorithm. However, the added utility to an adversary in this situation, as compared to what the adversary could have done without the requested table, was very small.[162] Whether the anticipated abuse is direct marketing or mindless harassment, the identification of *likely* neuroblastoma patients who are adduced from the purchased datasets will do the trick. Whether the hypothetical adversary is a pharmaceutical company or

157. *Id.* at 9 (alterations in original).

158. *Id.* at 20 (quoting S. Illinoisan v. Dep't of Pub. Health, 812 N.E.2d 27, 29 (Ill. App. Ct. 2004)).

159. *Id.* at 13.

160. *Id.*

161. *Id.* at 8.

162. More generally, the National Research Council has noted that in cases where "the same data are available elsewhere, even if not in the same form or variable combination, the added risk of releasing a research data file may be comparatively small." COMM. ON NAT'L STATISTICS, NAT'L RESEARCH COUNCIL, IMPROVING ACCESS TO AND CONFIDENTIALITY OF RESEARCH DATA 12 (Christopher Mackie & Norman Bradburn eds., 2000), *available at* http://www.geron.uga.edu/pdfs/BooksOnAging/ConfRes.pdf.

an Erin Brockovich-style environmental torts firm, the adversary could direct its solicitations to the set of likely candidates derived from the purchased, non-anonymized datasets. Dr. Sweeney testified that the requested cancer registry data was the "gold standard" that allowed her to re-identify the patients with confidence,[163] but this overstates the importance of the registry data tables since, without the government's verification, an attacker could still identify the likely candidates with enough confidence for her purposes.

Similarly, Narayanan and Shmatikov overstate the harm that can flow from re-identifying subjects in the Netflix database. Narayanan and Shmatikov explain that their algorithm works best when the movies reviewed on IMDb are less popular films.[164] The authors go into vivid detail in describing the movies that their two re-identified subjects rated in the Netflix database and draw absurd conclusions from them.[165] But they provide no information about the movies that the targets had freely chosen to rate publicly on IMDb using their real names — that is, the information that Narayanan and Shmatikov *already knew* before re-identifying them in the Netflix data. This information is crucial for understanding the marginal utility to putative adversaries. The inferences that are being drawn from the Netflix ratings — that they reveal political affiliation, sexual orientation, or, as the complaint for a recent lawsuit against Netflix alleges, "personal struggles with issues such as domestic violence, adultery, alcoholism, or substance abuse"[166] — can be drawn just as easily from the set of movies that the target had publicly rated in the first place. If the adversary already knows five or six movies that the target has watched, *that* knowledge can go a long way toward pigeonholing and making assumptions about the target.[167]

Of course, it is possible that a public data release could provide a great deal of extra information that would be valuable to a malfeasor.[168] But too often the marginal value is assumed to be very high

---

163. *S. Illinoisan*, 844 N.E.2d at 8.

164. *See* Narayanan & Shmatikov, *supra* note 11, at 116.

165. *Id.* at 123 ("[H]is political orientation may be revealed by his strong opinions about 'Power and Terror: Noam Chomsky in Our Times' and 'Fahrenheit 9/11,' and his religious views by his ratings on 'Jesus of Nazareth' and 'The Gospel of John.'").

166. Doe Complaint, *supra* note 5, at 18.

167. Privacy policy should not aspire to regulate these wrong-headed inferences; plenty of heterosexuals enjoyed Brokeback Mountain, and plenty of liberals dislike Michael Moore. But even if movie reviews are windows to the soul, the marginal information gained by re-identifying somebody in the Netflix dataset is likely to be small.

168. Education datasets often tie non-identifying but highly sensitive information (such as GPA or test scores) to indirect identifiers like age, race, and geography. If individuals in these databases were re-identified using the indirect identifiers, the adversary could learn something significant about the data subjects. *See, e.g.*, Krish Muralidhar & Rathindra Sarathy, *Privacy Violations in Accountability Data Released to the Public by State Educational Agencies*, FED. COMM. ON STAT. METHODOLOGY RES. CONF. 1 (Nov. 2009), http://www.fcsm.gov/09papers/Muralidhar_VI-A.pdf.

without any effort to compare the privacy risks after data release to the risks that exist irrespective of the data release.[169] More to the point, the accretion problem described by Paul Ohm — the prediction that increasing quantities of anonymized data will make re-identification of a rich data profile of us all the more possible[170] — is likely to be overshadowed by the accretion of *identified* data. Given the data mining opportunities available on identifiable information from companies like LexisNexis and Acxiom that aggregate identified information from private insurance and credit companies as well as public records,[171] it is highly unlikely that an adversary will find it worth his time to learn the Shannon entropy formula so that he can apply the Netflix algorithm.

IV. THE SKY IS NOT FALLING: THE REALISTIC RISKS OF PUBLIC DATA

The previous Part provided evidence that the focus of influential computer science literature is preternaturally consumed by hypothetical risks.[172] Unfortunately, legal scholars have taken up the refrain and have come to equally alarmist conclusions about the current state of data sharing.

In considering a public-use dataset's disclosure risk, data archivists focus on marginal risks — that is, the increase in risk of the disclosure of identifiable information compared to the pre-existing risks independent from the data release.[173] Just as the disclosure risk of a data release is never zero, the pre-existing risk to data subjects irrespective of the data release is also never zero. There are always other possible means for the protected information to become public unin-

169. Jeremy Albright, a researcher at the Interuniversity Consortium for Political and Social Research ("ICPSR"), notes that the statistical disclosure control literature has considered this approach but has generally not put it into practice, in part because nobody agrees on how much information the putative adversary should be presumed to have ahead of time. Jeremy Albright, Privacy Protection in Social Science Research: Possibilities and Impossibilities 11–12 (June 1, 2010) (unpublished manuscript) (on file with author).

170. Ohm, *supra* note 4, at 1746.

171. *See* ACXIOM CORP., UNDERSTANDING ACXIOM'S MARKETING PRODUCTS 1 (2010), *available at* http://www.acxiom.com/uploadedFiles/Content/About_Acxiom/Privacy/AC-1255-10%20Acxiom%20Marketing%20Products.pdf; *Risk Solutions Product Index*, LEXISNEXIS, http://www.lexisnexis.com/risk/solutions/product-index.aspx (last visited Dec. 21, 2011).

172. *See, e.g.*, Brickell & Shmatikov, *supra* note 128, at 70 (claiming that "[r]e-identification is a major privacy threat to public datasets containing individual records").

173. NAT'L RESEARCH COUNCIL, *supra* note 162, at 12. Thomas Louis of the University of Minnesota explains that disclosure risks associated with a particular data release should not be compared to a probability of zero, but that one should "consider how the probability of disclosure changes as a result of a specific data release." *Id.* Changes to the marginal risks caused by adding or masking certain fields in the dataset can be assessed as well. *Id.*

tentionally. How much marginal risk does a public research database create in comparison to the background risks we already endure?[174]

This Part assesses the realistic risks posed by the data commons. It lays out the frequency of improper anonymization and analyzes the likelihood that adversaries would choose re-identification as their means to access private information. The unavoidable conclusion is that contemporary privacy risks have little to do with anonymized research data.

A. Defective Anonymization

How often are public datasets released without proper anonymization? In other words, how often do data producers remove direct identifiers only, without taking the additional step of checking for subgroup sizes among indirect identifiers or without consideration to the discoverability of the sampling frame?

Paul Ohm discusses two high-profile examples: Massachusetts hospital data that failed to sufficiently cluster the indirect identifiers, and the AOL search query data that failed to remove last names.[175] This led two journalists at the New York Times to re-identify Thelma Arnold, who shared the spotlight with her search phrase "dog that urinates on everything."[176] Ohm argues that vulnerable public datasets with weak anonymization must be legion.[177] If sophisticated organizations like the Massachusetts Group Insurance Commission and AOL are not getting it right, what could we expect from a local agency?[178]

This concern has merit. A systematic study of disclosures made pursuant to the federal No Child Left Behind Act supports Ohm's in-

---

174. Releases of data by sophisticated data producers are expected, at a minimum, to "assure that the risk of disclosure from the released data when combined with other relevant publicly available data is very low." Working Paper No. 22, *supra* note 109, at 3. Of course, that begs the question what it means for disclosure risk to be "very low." Similarly, "[t]here can be no absolute safeguards against breaches of confidentiality, ... Many methods exist for *lessening* the likelihood of such breaches, the most common and potentially secure of which is anonymity." INT'L STAT. INST., *supra* note 14, at 10. Likewise, the FPCO's commentary on the newly passed FERPA regulations anticipate *low* risk, not the absence of risk altogether. "The regulations recognize that the risk of avoiding the disclosure of PII cannot be completely eliminated and is always a matter of analyzing and balancing risk so that the risk of disclosure is very low." FAMILY POLICY COMPLIANCE ORG., FAMILY EDUCATIONAL RIGHTS AND PRIVACY ACT, FINAL RULE, 34 CFR PART 99: SECTION-BY-SECTION ANALYSIS 11 (2008), *available at* http://www.ed.gov/policy/gen/guid/fpco/pdf/ht12-17-08-att.pdf.

175. Ohm, *supra* note 4, at 1717–20; *see also* Nate Anderson, *"Anonymized" Data Really Isn't — And Here's Why Not*, ARS TECHNICA (Sept. 8, 2009, 5:30 AM), http://arstechnica.com/tech-policy/news/2009/09/your-secrets-live-online-in-databases-of-ruin.ars.

176. Michael Barbaro & Tom Zeller Jr., *A Face is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES, Aug. 9, 2006, at A1.

177. Ohm, *supra* note 4, at 1729.

178. *Id.* at 1728.

tuition. The authors, Krish Muralidhar and Rathindra Sarathy, audited publicly available accountability data from several states to see whether the tabulations allow data users to glean PII.[179] While all of the states attempted to implement anonymization protocols, they all got it wrong one way or another.[180] Large repeat players in the data commons like the University of Michigan's Interuniversity Consortium of Policy and Social Research ("ICPSR") or the U.S. Census Bureau do not make these rookie mistakes, and often use data-swapping and noise-adding techniques for an additional level of security.[181] But the data commons no doubt contains some inadequately anonymized datasets that have not undergone best practices. This is almost certainly due to the abysmal state of the guidance provided by regulatory agencies and decisional law. There has not yet been a clear and theoretically sound pronouncement about the steps a data producer should take to reduce the risk of re-identification. I address this problem in Part V. For reasons I will elaborate on now, the risks imposed on data subjects by datasets that do go through adequate anonymization procedures are trivially small.

B. The Probability that Adversaries Exist

The "adversary" or "intruder" from the computer science literature is a mythical creature, the chimera of privacy policy. There is only a single known instance of de-anonymization for a purpose other than the demonstration of privacy risk,[182] and no known instances of a re-identification for the purpose of exploiting or humiliating the data subject. The Census Bureau has not had any known instances of data abuse, nor has the National Center for Education Statistics.[183]

This is not surprising, because the marginal value of the information in a public dataset is usually too low to justify the effort for an intruder. The quantity of information available in the data commons is outpaced by the growth in information self-publicized on the Internet or collected for commercially available consumer data. Consumer

---

179. Muralidhar & Sarathy, *supra* note 168, at 1.

180. *Id.* at 20.

181. RICHARD A. MOORE, JR., U.S. BUREAU OF THE CENSUS, CONTROLLED DATA-SWAPPING TECHNIQUES FOR MASKING PUBLIC USE MICRODATA SETS 25–26, *available at* http://www.census.gov/srd/papers/pdf/rr96-4.pdf.

182. Duff Wilson*, Database on Doctor Discipline is Restored, with Restrictions*, N.Y. TIMES, Nov. 10, 2011, at B2 (News organizations linked identifiable court filings to a national databank of doctor disciplinary actions in order to criticize the disciplinary boards. The journalists re-identified doctors who had a known, long history of malpractice actions against them to the "de-identified" data on disciplinary actions. The public-use data employed trivial anonymization — the removal of names only.).

183. *See* NAT'L RESEARCH COUNCIL, *supra* note 162, at 48; Hermann Habermann, *Ethics, Confidentiality, and Data Dissemination,* 22 J. OF OFFICIAL STAT. 599, 603 (2006).

data catalogs boast that businesses can "choose [an] audience by their ailments & medications."[184]

Unfortunately, privacy advocates routinely fail to report the dearth of known re-identification attacks.[185] Instead, scenarios of re-identification and public humiliation are held up like Desdemona's handkerchief, inspiring suspicion and fear for which we have, as yet, no evidence. As Paul Ohm says,

> Almost every person in the developed world can be linked to at least one fact in a computer database that an adversary could use for blackmail, discrimination, harassment, or financial or identity theft. I mean more than mere embarrassment or inconvenience; I mean legally cognizable harm. Perhaps it is a fact about past conduct, health, or family shame. For almost every one of us, then, we can assume a hypothetical database of ruin, the one containing this fact but until now splintered across dozens of databases on computers around the world, and thus disconnected from our identity. Reidentification has formed the database of ruin and given our worst enemies access to it.[186]

Ohm speaks in the present tense; he suggests the database of ruin has arrived.

It is possible that intruders are keeping their operations clandestine, reverse-engineering our public datasets without detection. But this conviction should not be embraced too quickly. Other forms of data-privacy abuse that ought to be difficult to detect have nevertheless come to light due to whistleblowing and sleuthing.[187] Paul Syver-

---

184. SPECIALISTS MKTG. SERVS., INC., MAILING LIST CATALOG, *available at* http://directdatamailinglists.com/SMS-catalog.pdf.

185. *See, e.g.*, DANIEL SOLOVE, THE DIGITAL PERSON 82–83, 173–74 (2008), *available at* http://docs.law.gwu.edu/facweb/dsolove/Digital-Person/text.htm; Ohm, *supra* note 4, at 1729; Brickell & Shmatikov, *supra* note 128, at 70 (claiming that "[r]e-identification is a major privacy threat to public datasets containing individual records"). Thomas M. Lenard and Paul H. Rubin notice this phenomenon, observing that while Solove's study "lists harms associated with information use, he does not quantify how frequent or serious they are." THOMAS M. LENARD & PAUL H. RUBIN, TECH. POLICY INST., IN DEFENSE OF DATA: INFORMATION AND THE COSTS OF PRIVACY 43 (2009), http://www.techpolicyinstitute.org/files/in%20defense%20of%20data.pdf.

186. Ohm, *supra* note 4, at 1748.

187. Pharmatrak, Inc. collected personally identifiable data on web visitors to its pharmaceutical industry clients using clear GIFs (or "cookies") in direct contravention of the Electronic Communications Privacy Act. This practice was exposed and resulted in a class action lawsuit. In re Pharmatrak, Inc., 329 F.3d 9, 12 (1st Cir. 2003). HBGary Federal considered hacking into the networks of its clients' foes in order to gather evidence for smear campaigns, but these practices were uncovered, ironically enough, during a hack into their

son suggests that we could test the hypothesis of covert re-identification by comparing the incidence of identity theft to behaviors or characteristics in accessible datasets to see if there is a correlation that might suggest these data subjects were re-identified at some point.[188] This experiment is worthwhile, but the available aggregate data suggests there is no such relationship. Identity theft plateaued between 2003 and 2009 and dropped to its lowest recorded level in 2010.[189] Moreover, the largest category of identity fraud schemes involves "friendly fraud" — fraudulent impersonation committed by people that know the victim personally (such as a roommate or relative) — and this category has grown in proportion while the other categories declined.[190] These statistics contradict the position that we are inching ever closer to our digital ruination.

Like any default hypothesis, the best starting point for privacy policy is to assume that re-identification does not happen until we have evidence that it does. Because there is lower-hanging fruit for the identity thief and the behavioral marketer — blog posts to be scraped and consumer databases to be purchased — the thought that these personae non gratae are performing sophisticated de-anonymization algorithms is implausible.

C. Scale of the Risk of Re-Identification in Comparison to Other Tolerated Risks

Privacy risks are difficult to measure and understand — to feel at a gut level.[191] One useful heuristic for comprehending the privacy risks of public anonymized data is to compare those risks to other privacy risks that we know and tolerate.

Our trash is a rich and highly accessible source of private information about us — indeed, it continues to have the distinction of being a tremendously valuable resource for private investigators and

---

own servers. *See* Eric Lipton & Charlie Savage, *Hackers' Clash with Security Firm Spotlights Inquiries to Discredit Rivals*, N.Y. TIMES, Feb. 11, 2011, at A15.

188. Paul Syverson, *The Paradoxical Value of Privacy*, 2D ANN. WORKSHOP ON ECON. & INFO. SECURITY 2 (2003), http://www.cpppe.umd.edu/rhsmith3/papers/Final_session3_syverson.pdf.

189. *The Notable Decline of Identity Fraud*, HELP NET SECURITY (Feb. 8, 2011), www.net-security.org/secworld.php?id=10551; *see also* LENARD & RUBIN, *supra* note 185, at 34–35. The aggregate data cannot directly answer the question about the relationship between public data and identity theft. Ironically, microdata is required to reliably test this theory of covert re-identification.

190. *See The Notable Decline of Identity Fraud*, *supra* note 189.

191. This is at the heart of Peter Swire's criticism of scholars like me who attempt to compare the costs and benefits of privacy. *See* Peter Swire, Privacy and the Use of Cost/Benefit Analysis 4, 10 (June 18, 2003) (unpublished manuscript), *available at* http://www.ftc.gov/bcp/workshops/infoflows/present/swire.pdf.

identity thieves.[192] Data presents no more risk (and often less risk) than our garbage. Thomas Lenard and Paul Rubin have noted that breach notification requirements and other warnings about the privacy hazards of conducting business online could lead consumers to conduct business offline and demand paper statements. Ironically, this result would greatly increase the likelihood of identity theft.[193]

Moreover, consider the large quantity of sensitive personally identifiable information available in public records. Income information, thought to be among the most sensitive categories of information,[194] is available for most public employees.[195] The names and salaries of the highest-paid employees in California are tracked on the Sacramento Bee's website.[196] Litigants and witnesses in lawsuits are often forced to divulge personal information and face embarrassing accusations, and juror identities and questionnaire responses are usually within the public domain.[197] We accept these types of exposures because the countervailing interests — ensuring transparency and accountability in state action — warrant it. The Constitution protects these types of disclosures through a robust set of First Amendment precedents, and the tradeoffs in terms of privacy invasions have proven to be bearable to society.[198]

The closest cousin to the malicious de-anonymizer is the hacker. This type of adversary certainly exists. If we are to imagine a skilled computer programmer determined to find out a target's secrets, is it not easier to imagine him just hacking into the target's personal computer? This, after all, was HBGary Federal's modus operandi when it consulted to do the dirty work for Bank of America, corporate law firms, and their clients.[199] HBGary Federal planned to create extensive dossiers of rivals or critics for the purpose of forming smear campaigns.[200] When HBGary Federal proposed to make a dossier on members of U.S. Chamber Watch, a consumer watchdog organiza-

---

192. Frank Abagnale stresses the importance of eliminating the garbage and paper trail to reduce the risk of identity fraud. *See Abagnale Recommends Fraud Protection Strategy: Audio*, BLOOMBERG (Nov. 15, 2010), http://www.bloomberg.com/news/2010-11-15/abagnale-recommends-fraud-protection-strategy-audio.html.

193. LENARD & RUBIN, *supra* note 185, at 38–39.

194. *See* Bernardo A. Huberman, Eytan Adar & Leslie R. Fine, *Valuating Privacy*, IEEE SECURITY & PRIVACY, Sept.–Oct. 2005, at 22, 22–24.

195. For example, see the salary information available online for University of Michigan employees at http://www.umsalary.info/deptsearch.php.

196. *State Worker Salary Search: Top Salaries Earned in 2010*, THE SACRAMENTO BEE, http://www.sacbee.com/statepay (last visited Dec. 21, 2011); *see also* Comm'n on Peace Officer Standards and Training v. Superior Court, 165 P.3d 462, 465 (Cal. 2007).

197. *See, e.g.*, Pantos v. City & Cnty. of S.F., 198 Cal. Rptr. 489, 491 (Cal. Ct. App. 1984); Forum Commc'ns Co. v. Paulson, 752 N.W.2d 177, 185 (N.D. 2008).

198. *See* Cox Broad. Corp. v. Cohn, 420 U.S. 469, 496 (1975); Fla. Star v. B.J.F., 491 U.S. 524, 538 (1989).

199. Lipton & Savage, *supra* note 187.

200. *Id.*

tion, their plans included identifying vulnerabilities in the targets' computer networks that could be exploited.[201] HBGary Federal responded to the incentives to engage in unethical and illegal behavior to garner the favor of its clients. Yet, it is difficult to imagine that HBGary's agenda would ever include re-identifying their targets in public-use anonymized datasets. The alternative approaches are so much easier.

A malfeasor with no specific target in mind is still better off using hacking techniques rather than de-anonymization algorithms. That is what hackers did to expose 236,000 mammography patient records at the University of North Carolina School of Medicine,[202] 160,000 health records for University of California students,[203] and 8,000,000 records in the Virginia Prescription Monitoring Program (for which the hackers sought a $10 million ransom).[204] These sorts of hacks require significantly less skill than the de-anonymization of a research dataset because malware capable of exploiting bugs in popular programs and operating systems is sold on the black market to whomever is unethical enough to use it.[205] The programs require little to no customization because they apply malicious code to popular programs that all suffer from identical vulnerabilities.[206] De-anonymization algorithms, in contrast, require a theoretical understanding of the algorithm in order to suit the attack to a particular dataset.[207]

Data spills — the mishandling of unencrypted data — provide another illustration of the risk of re-identification. These spills typically expose the personally identifiable information of customers or patients. In the last couple years the medical records of 7.8 million people have been exposed in various sorts of security breaches.[208] The

---

201. *Id.* Ironically, it was HBGary Federal's own networks' vulnerabilities that it should have been focusing on, as the hacker group Anonymous hacked into HBGary Federal's servers and released several emails and PowerPoint presentations on Wikileaks. *Id.*

202. *Hackers Attack UNC-Based Mammography Database,* UNC HEALTH CARE (Sept. 25, 2009), http://news.unchealthcare.org/som-vital-signs/archives/vital-signs-sept-25-2009/hackers-attack-unc-based-mammography-database.

203. *Hackers Get Into U.C. Berkeley Health-Records Database*, FOXNEWS.COM (May 8, 2009), http://www.foxnews.com/story/0,2933,519550,00.html.

204. Brian Krebs, *Hackers Break into Virginia Health Professions Database, Demand Ransom*, WASH. POST SECURITY FIX (May 4, 2009, 6:39 PM), http://voices.washingtonpost.com/securityfix/2009/05/hackers_break_into_virginia_he.html.

205. *See* Derek E. Bambauer & Oliver Day, *The Hacker's Aegis*, 60 EMORY L.J. 1051, 1101 (2011); *see also* Larry Barrett, *Data Theft Trojans, Black Market Cybercrime Tools on the Rise*, ESECURITY PLANET (Mar. 31, 2010), http://www.esecurityplanet.com/trends/article.php/3873891/Data-Theft-Trojans-Black-Market-Cybercrime-Tools-on-the-Rise.htm.

206. *See* Bambauer & Day, *supra* note 205, at 1060–62; Jaziar Radianti & Jose J. Gonzalez, Toward a Dynamic Modeling of the Vulnerability Black Market 4–7 (Oct. 23–24, 2006) (unpublished manuscript), *available at* http://wesii.econinfosec.org/draft.php?paper_id=44.

207. *See supra* Part III.

208. Milt Freudenheim, *A New Push to Protect Health Data*, N.Y. TIMES, May 31, 2011, at B1.

spills are often the result of improper handling by employees who were authorized to access the information. For example, Massachusetts General Hospital recently agreed to pay a one million dollar fine after one of its employees lost the records of 192 patients on the subway, many of whom had HIV/AIDS.[209] So the question for our purposes is this: if we are to fear users of public anonymized datasets, why do we tolerate the handling of our personal information by minimally paid, unskilled data processors?[210] (Indeed, some companies have used prison labor to perform data entry.[211])

The intuitive answer is that data has become the lifeblood of our economy. It is more rational to spread risk among all the consumers and modify data handling behavior through fines and sanctions than it is to expect consumers to forego the convenience and customized service of the information economy.[212] It is puzzling, then, why privacy advocates have chosen to target anonymized research data — data that poses relatively low risk to the citizenry and offers valuable public-interest-motivated research in return — as a cause worthy of preemptive strike.[213]

V. A PROPOSAL IN THE STATE OF HIGHLY UNLIKELY RISK

The fractured set of privacy statutes and rules in the United States generally requires data producers to refrain from releasing data that can be used to re-identify a data subject.[214] A great limitation of current U.S. privacy law — a limitation that runs against the interests of the data subjects and researchers alike — is that privacy law regulates

209. *See Morning Edition: MGH Settles for $1M over Lost HIV/AIDS Records*, NAT'L PUB. RADIO (Feb. 25, 2011), http://www.wbur.org/2011/02/25/mgh-privacy.

210. In Massachusetts General Hospital's case, the employee was a billing manager, and not a low skilled employee. *Id.* But records, particularly consumer records, are often in the hands of low skill data processors or outsourced to third parties that process the data offshore. *See, e.g.*, *Outsourcing Data Entry Privacy Policy*, DATA ENTRY SERVICES INDIA, http://www.dataentryservices.co.in/privacy_policy.htm (last visited Dec. 21, 2011). Not everybody is comfortable with the risk that accompanies routine data-handling. Parents of a student who participated in a research survey at their child's school attempted to mount a legal challenge based on the potential privacy risks that an administrator might divulge their child's information inadvertently, but the suit was dismissed. C.N. v. Ridgewood Bd. of Educ., 430 F.3d 159, 161 (3d Cir. 2005).

211. Sandra T.M. Chong, *Data Privacy: The Use of Prisoners for Processing Personal Information*, 32 U.C. DAVIS L. REV. 201, 204 (1998).

212. *See* Cate, *supra* note 9, at 12–16. Poor encryption practices are an excellent target for effective privacy regulation. There is no reason for a business or agency to fail to encrypt its files that contain personally identifiable information. *See* Derek E. Bambauer, *Rules, Standards, and Geeks*, 5 BROOK. J. CORP. FIN. & COM. L. 49, 56–57 (2010).

213. A wiser target is the law surrounding data security breaches. *See, e.g.*, Paul M. Schwartz and Edward J. Janger, *Notification of Data Security Breaches*, 105 MICH. L. REV. 913 (2007); *see also* Bambauer, *supra* note 212, at 49.

214. *See supra* text accompanying notes 20–21.

the *release* of data rather than its use.[215] Privacy law does not prohibit an end-user from re-identifying somebody in a public-use dataset. Rather, the laws and statutory schemes act exclusively on the releaser.[216] In many respects, the current approach to data privacy is dissatisfying to the full range of affected parties, and we are beginning to see an influx of new proposals.

The most popular suggestions for altering data privacy laws differ in their particulars, but they invariably impose large transaction costs on research, if they do not preclude it altogether. The FTC's recently unveiled framework for consumer data advises companies not to distinguish between anonymized and personally identifiable data, which means that anonymized research data must be subjected to the exact same limitations imposed on the collection and use of identifiable data.[217] This vision bars private companies from participating in the data commons, since a public release of research data would be treated the same as a security breach or a spill of identifiable data. The FTC's framework borrows from the European Data Protection Directive, which requires the unambiguous consent of data subjects before personal data can be processed into statistical research data.[218] If the FTC framework is a harbinger for what is to come, the data commons is in real trouble.[219]

Paul Ohm and Daniel Solove propose "contextual" privacy regulations to bring legal liability in line with the risk that the data producer has created.[220] Ohm suggests that a data releaser should consider all the determinants of re-identification risk and assess whether a threat to the data subject exists.[221] While this solution has natural appeal as a levelheaded approach, a loose case-by-case standard will provide little guidance and assurance for data producers. In fact, existing statutes already implement the bulk of the suggestions Ohm puts forward. HIPAA regulations, for example, instruct data producers to remove

---

215. *Id.*

216. *Id.*

217. FTC PRIVACY REPORT, *supra* note 5, at 43, 51–52.

218. *See* Directive 95/46/EC, of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 2001 O.J. (L 281) 31, 34, 40. Note that under Recital 29, processing for statistical purposes is, at least, not a use inconsistent with any other use for which the data may be processed. *Id.* at 34. Social science researchers often have to perform their analyses at the physical location of the data enclaves. *See, e.g.*, Stefan Bender et al., *Improvement of Access to Data Set from the Official Statistics* 4–5 (German Council for Soc. and Econ. Data, Working Paper No. 118, 2009), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1462086.

219. Legislation seeking to limit the storage of data has already been proposed. *See, e.g.*, Eliminate Warehousing of Consumer Internet Data Act of 2006, H.R. 4731, 109th Cong. (2006).

220. Ohm, *supra* note 4, at 1762; Daniel J. Solove, *Conceptualizing Privacy*, 90 CALIF. L. REV. 1087, 1091–93 (2002).

221. Ohm, *supra* note 4, at 1764.

any information that, in context, might lead to the re-identification of a data subject, and they differentially scrutinize public releases much more severely, while giving agencies and firms wide latitude when drawing up licenses with business associates.[222] But for the reasons detailed in Part II, these standards are encouraging over-protectionism and providing agencies with an evasion tactic. Moreover, licensing processes impose transaction costs on researchers that are not justified by the speculative risks of re-identification.

I propose something altogether different: simple, easy-to-apply rules.[223] My policy has three aspects to its design: (1) it clarifies what a data producer is expected to do in order to anonymize a dataset and avoid the dissemination of legally cognizable PII; (2) it immunizes the data producer from privacy-related liability if the anonymization protocols are properly implemented; and (3) it punishes with harsh criminal penalties any recipient of anonymized data who re-identifies a subject in the dataset for an improper purpose. I will describe each of these aspects in more detail and explain why the proposed approach offers an improvement over current laws and regulations.

A. Anonymizing Data

Under my approach, a data producer is required to do just two things in order to convert personally identifiable data into anonymized (non-PII) data: (1) strip all direct identifiers, and (2) either check for minimum subgroup sizes on a preset list of common indirect identifiers — such as race, sex, geographic indicators, and other indirect identifiers commonly found in public records — or use an effective random sampling frame.

(1) *Stripping Direct Identifiers.* The removal of direct identifiers (name, telephone number, address, social security number, IP addresses, biometric identifiers like fingerprints, and any other unique identifying descriptor) is an obvious first step, but one that should not go without comment. After all, this critical oversight led to the re-

222. For instance, under HIPAA, the public release of health information requires the covered entity to prepare the data such that "there is no reasonable basis to believe that the information can be used to identify an individual." 45 C.F.R. § 164.514(a) (2010). Releases of identifiable health information to a business associate, on the other hand, are permitted so long as the business associate makes assurances that it will guard and handle the health data in a manner consistent with the covered entity's responsibilities under HIPAA. *Id.* § 164.502(e)(1)(i) (2010). Any additional restrictions the covered entity might wish to impose are left to the original data-holder's discretion.

223. Derek Bambauer argues that rules are more helpful than standards in contexts when three conditions are met: (1) when the specified minimum standard for behavior will suffice most or all of the time, (2) when the standard degrades slowly, and (3) when monitoring for harm is low-cost and accurate. Bambauer, *supra* note 212, at 50. Here, the first condition is met because, as I argued earlier, re-identificaiton attacks performed on anonymized data are difficult, and anonymization has sufficed to prevent re-identification attacks. *See supra* Parts III, IV. The second and third conditions are developed in this Part.

identification of a data subject in the AOL search term database.[224] Remarkably, the privacy community and even the FTC have held this up as a key exemplar for the proposition that there is no viable way to adequately anonymize data anymore.[225] In fact, the AOL story is an example of a *lack* of anonymization.

(2) *Basic Risk Assessment.* My next step requires the data producer either to count the *minimum subgroup sizes* or to confirm that the dataset has an *unknown sampling frame*. Neither of these is conceptually difficult.

Minimum Subgroup Count — This ensures that no combination of indirect identifiers yields fewer than a certain threshold number of observations (usually between three and ten). For the purpose of this Article I will use five.[226] This is known as "k-anonymity" in the computer science literature.[227] Suppose a college wishes to release a public-use version of its grades database. If there are only two Asian female chemistry majors in the cohort of students that entered in 2010, then the school should not release a dataset that includes race, gender, major, and cohort year unless it first blurs together some of these categories. The college might choose to lump several majors together into clusters, or lump cohort years into bands spanning five years. There are a number of ways to blur the categories such that minimum subgroup counts stay above the required threshold. Indirect identifiers are limited to categories of information that are publicly available for all or most of the data subjects — e.g., age, gender, race, and geographic location. They do not include information that is not systematically compiled and distributed by third parties.[228]

Unknown Sampling Frame — If a public data user has no basis for knowing whether an individual is in the universe of people described in the dataset, then the dataset does not — and cannot — disclose PII. Sampling frame is a powerful tool for anonymizing data, and large statistical bureaus (such as the U.S. Census Bureau) often employ it when they collect information on a random sample of

224. AOL failed to strip the dataset of last names. This oversight, in combination with multiple searches for a particular neighborhood, led to the re-identification of Thelma Arnold. Barbaro & Zeller, *supra* note 176.

225. FTC PRIVACY REPORT, *supra* note 5, at 36, 38. The AOL story, along with the Netflix study, was the support for the FTC's broad-reaching conclusion that "businesses combine disparate bits of 'anonymous' consumer data from numerous different online and offline sources into profiles that can be linked to a specific person." *Id.*

226. The Centers for Disease Control and Prevention anticipates aggregated tables using a threshold value of three. CTRS. FOR DISEASE CONTROL AND PREVENTION & HEALTH RES. SERVS. ADMIN., INTEGRATED GUIDELINES FOR DEVELOPING EPIDEMIOLOGIC PROFILES 126 (2004), *available at* http://www.cdc.gov/hiv/topics/surveillance/resources/guidelines/epi-guideline/pdf/epi_guidelines.pdf.

227. *See* Sweeney, *supra* note 18, at 557.

228. For example consumer preferences and information contained on a Facebook "wall" are not indirect identifiers in my scheme.

Americans.[229] Thus, if the Bureau of Labor Statistics produces a dataset that includes only one veterinarian in Delaware, we need not be concerned unless there is some way to know which of the many veterinarians in Delaware the dataset is describing. If the sampling frame is unknown, then the minimum subgroup count and extremity-coding rules need not apply.[230] But precautions must be taken to ensure that an outsider really cannot discern whether the sample includes a particular individual.[231]

If either of these protocols is properly implemented, the dataset would be legally recognized as anonymized non-PII data. To be clear, this standard is *less* onerous than the current state and federal laws like HIPAA. This is by design. While my proposal diverges sharply from others', it flows naturally from the assertion, supported earlier in this Article, that the risk of re-identification is not significant. Nevertheless, agencies and organizations that work with data frequently enough to have Institutional Review Boards should continue to use heightened standards determined by current best practices.[232] The procedures described above set an appropriate floor, and need not be interpreted as a ceiling.

Freeing up the flow of data will enrich the proverbial marketplace of ideas. In the past, the simplified process of stripping obvious identifiers was legally sufficient to protect an individual's privacy.[233] We have drifted into protecting against more and more intricate attacks without having experienced any of them. Moreover, some of the more complex disclosure-risk avoidance techniques (such as data-swapping or noise-adding) have gone awry. The U.S. Census Bureau's public-use microdata samples ("PUMS files") from the 2000 census contain

---

229. For example, the Public-Use Microdata Samples ("PUMS files") report data on a sample of U.S. households. *See Public-Use Microdata Samples (PUMS)*, U.S. CENSUS BUREAU, http://www.census.gov/main/www/pums.html (last updated May 28, 2010).

230. This assumption can fail in circumstances where a potential data subject is unusual. If the indirect identifiers included in the dataset uniquely describe a person in the broad population of people that could potentially be included in the sample, an adversary will be able to check whether that person actually *is* included in the sample (and identify him if he is). For example, suppose only one veterinarian in Delaware identifies himself as a Native American; a dataset that included profession, state, and detailed race information cannot rely on an unknown sampling frame to ensure anonymity because any dataset including these indirect identifiers would immediately identify the individual in question as being a member of the dataset.

231. *See, e.g.*, Khaled El Emam & Fida Kamal Dankar, *Protecting Privacy Using k-Anonymity*, 15 J. AM. MED. INFO. ASS'N 627, 634–35 (2008).

232. *See* George T. Duncan, *Confidentiality and Data Access Issues for Institutional Review Boards*, *in* PROTECTING PARTICIPANTS AND FACILITATING SOCIAL AND BEHAVIORAL SCIENCES RESEARCH 235, 235 (Constance F. Citro et al. eds., 2003).

233. *See* Nat'l Cable Television Ass'n v. FCC, 479 F.2d 183, 195 (D.C. Cir. 1973); Tax Analysts and Advocates v. IRS, 362 F. Supp. 1298, 1307 (D.D.C. 1973) (quoting *Nat'l Cable Television Ass'n*, 479 F.2d at 195)

substantial errors in the reporting of age and gender that have affected analyses for a decade's worth of research.[234]

B. Safe Harbor for Anonymized Data

If a data producer follows the anonymization protocols, it will be shielded from liability based on privacy torts, certain types of contractual liability, and federal statutory penalties defined by privacy statutes like HIPAA. The anonymization protocols would also take the data out of the ambit of privacy exemptions in public records statutes (meaning that government agencies legally obligated to disclose information through public records laws could not make use of the privacy exemption if a useful dataset could be produced using the anonymization procedures described above). With the exception of contractual liability, on which I elaborate below, the scope of this safe harbor provision is fairly predictable.

The safe harbor provision protects data producers from liability based on confidentiality agreements unless the confidentiality agreement explicitly prohibits the dissemination of all information, whether or not it is in identifiable form, to any unnamed third parties. To be clear, if the firm collecting data reserves the right to share information to a third party in the private agreement, anonymized data will not violate the confidentiality agreement. The reason for structuring the safe harbor provision this way is to prevent the very likely scenario in which a company wishes to profit from the information it collects by sharing it with marketers or business partners, while simultaneously having a consumer-friendly-sounding excuse for shielding anonymized data from researchers who might use the data to uncover fraud or discrimination. Of course, nothing in this scheme obligates an organization to share anonymized research data, but it does remove the fig leaf — the pretense of sensitivity — when data is shared for marketing and business purposes.

Immunity is bold, but it is not unusual for the law to go to great lengths to bolster the public's interest in information. Courts have been especially protective of the First Amendment right to disseminate truthful information of public concern.[235] In the context of undercover journalism, scholars and lawmakers have concluded that the public interest in unearthing information justifies immunity from tort liability, even when journalists employ deceptive newsgathering prac-

---

234. *See* J. Trent Alexander, Michael Davern & Betsey Stevenson, *Inaccurate Age and Sex Data in Census PUMS Files* 1–3 (CESifo Working Paper No. 2929, 2010), *available at* http://ssrn.com/abstract=1546969; Steven Levitt, *Can You Trust Census Data?*, FREAKONOMICS (Feb. 2, 2010, 11:09 AM), http://www.freakonomics.com/2010/02/02/can-you-trust-census-data.

235. *See* Bartnicki v. Vopper, 532 U.S. 514, 515, 518 (2001); Fla. Star v. B.J.F., 491 U.S. 524, 525 (1989); Sidis v. F-R Publ'g Corp., 113 F.2d 806, 807–09 (2d Cir. 1940).

tices.[236] C. Thomas Dienes notes that "[i]n the private sector, when the government fails in its responsibility to protect the public against fraudulent and unethical business and professional practices, whether because of lack of resources or unwillingness, media exposure of such practices can and often does provide the spur forcing government action."[237] Likewise, Erwin Chemerinsky defends paparazzi-style journalism by reminding the academy:

> Speech is protected because it matters in people's lives, and aggressive newsgathering is often crucial to obtaining the information. The very notion of a marketplace of ideas rests on the availability of information. . . . People on their own cannot expose unhealthy practices in supermarkets or fraud by telemarketers or unnecessary surgery by doctors. But the media can expose this, if it is allowed the tools to do so, and the public directly benefits from the reporting.[238]

Undeniably, the data commons is one of these tools. It provides invaluable probative power that cannot be matched by anecdote or concentrated theorizing, and the risk of re-identification is relatively small compared to the informational value.

C. Criminal Penalties for Data Abuse

Finally, the safe harbor must be buttressed by a statute that criminalizes and stiffly punishes the improper re-identification of subjects within a properly anonymized dataset.[239] Criminal liability attaches the instant an adversary discloses the identity and a piece of non-public information to one other person who is not the data producer.[240] First, this design avoids unintentionally criminalizing disclosure-risk research — research that can usefully identify vulnerabilities in anonymized datasets. This sort of information will be invaluable to

---

236. *See* Desnick v. ABC, 44 F.3d 1345, 1354–55 (7th Cir. 1995); Erwin Chemerinsky, *Protect the Press: A First Amendment Standard for Safeguarding Aggressive Newsgathering*, 33 U. RICH. L. REV. 1143, 1160 (2000). *But see* Food Lion, Inc. v. ABC, 194 F.3d 505, 521 (4th Cir. 1999).

237. C. Thomas Dienes, *Protecting Investigative Journalism*, 67 GEO. WASH. L. REV. 1139, 1143 (1999).

238. Chemerinsky, *supra* note 236, at 1160.

239. In order to trigger criminal protection against re-identification, the dataset must be properly anonymized in accordance with the requirements affording safe harbor protection. This prevents users of a poorly anonymized dataset from incurring criminal liability.

240. If the sample frame is unknown, non-public information can consist of information reported about the subject in the dataset or even the mere fact that the subject is in the dataset.

data producers and regulators if an attack seems likely to be replicated by a true malfeasor. De-anonymization scientists will be able to continue publishing their work with impunity. Second, this design avoids the possibility of innocent technical violations by requiring an overt, malicious act — disclosing a non-public piece of information to one other person.[241]

Current privacy statutes leave a blatant gap in coverage: they do not restrain an adversary from re-identifying a subject. To address this, the Institute of Medicine of the National Academies has proposed legal sanctions for re-identification,[242] and Robert Gellman has proposed a system of data sharing through uniform licensing agreements that protect against the re-identification of data subjects using criminal and civil sanctions.[243] In fact, much of the public research data available to researchers today requires the execution of data license agreements prohibiting re-identification and requiring the research staff to ensure the security of the data.[244] A federal criminal statute would provide uniform protection for all data subjects, and would reduce transaction costs between data users and data producers by making contractual promises of this sort unnecessary.

The criminal penalty is particularly important when a dataset has been properly anonymized, but an adversary decides to target a specific data subject about whom the adversary has special information. Take the following example, which comes from the Department of Education's commentary on the 2009 revisions of the FERPA regulations:

> [I]f it is generally known in the school community that a particular student is HIV-positive . . . then the school could not reveal that the only HIV-positive student in the school was suspended. However, if it

241. I do not believe this precaution is necessary to avoid "thought crimes." After all, tort law has made actionable some forms of observation in public. For example, even mere public surveillance can be actionable under the tort of intrusion. *See* Summers v. Bailey, 55 F.3d 1564, 1566 (11th Cir. 1995); Nader v. Gen. Motors Corp., 255 N.E.2d 765, 769–71 (N.Y. 1970). The reverse-engineering of an anonymized dataset is at least as intrusive and requires just as much actus reus.

242. INST. OF MED. OF THE NAT'L ACADS., BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH 265 (Sharyl J. Nass et al. eds., 2009), *available at* http://www.ncbi.nlm.nih.gov/books/NBK9578/pdf/TOC.pdf.

243. Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 33, 51–52 (2010). Paul Ohm also suggests that regulators should consider prescribing "new sanctions — possibly even criminal punishment — for those who reidentify." Ohm, *supra* note 4, at 1770. Both Gellman's and the Institute of Medicine's proposals restrict researchers from sharing the de-identified data outside their research teams. Gellman, *supra*, at 51–52; INST. OF MED. OF THE NAT'L ACADS., *supra* note 242, at 49–50. My proposal does not prohibit re-disclosure of anonymized data.

244. *See, e.g.*, *Restricted Data Use Agreement*, ICPSR, http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/restricted/agreement.jsp (last visited Dec. 21, 2011).

> is not generally known or obvious that there is an HIV-positive student in school, then the same information could be released, even though someone with special knowledge of the student's status as HIV-positive would be able to identify the student and learn that he or she had been suspended.[245]

Likewise, someone with special knowledge about the circumstances of a particular student's suspension could use that information to discern that he or she is HIV-positive. While the student might have civil recourse if the adversary publicizes this fact and causes sufficient harm,[246] nothing in FERPA's design outlaws the adversary's acts in re-identifying the student in the first place. The heavy hand of the prosecutor is an appropriate means for enforcing the ethics of the data commons.

Though detection and enforcement of this provision would no doubt be very difficult, this does not mean that retributive disincentives have no effect. People and firms often overreact to improbable but unknown risks of criminal sanction.[247] Moreover, one major motivation for my proposal is the understanding that re-identification is unlikely to happen. Thus, the criminal element to this data privacy scheme is, by design, expensive and likely to operate more as a disincentive than as a penalty actually imposed by courts.

D. Objections

The objection to my framework is simple: What if I am wrong? By the time we realize that anonymization can be undone, it is too late! Ohm's contention is that data that cannot re-identify us today will be capable of doing so tomorrow.[248] We need urgent action because we are laying the groundwork for the "database of ruin."[249] This argument shares a remarkable resemblance to fears about the introduction of computers into the federal government in the 1960s. The statement of Representative Cornelius E. Gallagher of New Jersey before the Committee on Government Operations is typical of these fears:

---

245. Family Educational Rights and Privacy, 73 Fed. Reg. 74,806, 74,832 (Dec. 9, 2008).

246. The student might be able to bring a claim based on the tort of public disclosure of private facts. *See* RESTATEMENT (SECOND) OF TORTS § 652D (1977).

247. *See* John E. Calfee & Richard Craswell, *Some Effects of Uncertainty on Compliance with Legal Standards*, 70 VA. L. REV. 965, 966 (1984). This reaction is also reflected in the high prices firms pay for criminal liability insurance. *See* Miriam H. Baer, *Insuring Corporate Crime*, 83 IND. L.J. 1035, 1036 (2008).

248. Ohm, *supra* note 4, at 1748, 1757.

249. *Id.* at 1757.

> Nor do we wish to see a composite picture of an individual recorded in a single informational warehouse, where the touch of a button would assemble all the governmental information about the person since his birth. . . . Although the personal data bank apparently has not been proposed as yet, many people view this proposal as a first step toward its creation. . . . We cannot be certain that such dossiers would always be used by benevolent people for benevolent purposes.[250]

Anxieties over potential abuse of new information technologies are a hardy perennial.[251] Today, the threatening technology is the Internet. While the Internet certainly increases the risk of re-identification, and while producers of anonymized data should be cognizant of new and rich collections of auxiliary information available to a malicious intruder, the additional risk is not as great as it might seem. Remember that, in order to re-identify a subject in a dataset, an adversary must be confident that a unique data subject matches a unique member of the general population.[252] Suppose an anonymized prescription dataset described a fifty-year-old woman in central Vermont who is taking pharmaceutical drugs to treat depression and high cholesterol. An adversary comes across a LiveJournal blog post by a woman who identifies herself, reveals that she is fifty years old and living in Montpelier, and describes her experience on Lipitor.[253] The adversary has stumbled upon a likely candidate to match up to the anonymized data subject, and if he is right, he will have learned that the blogger is also clinically depressed. But in order to be confident in the match, he must have some reason to believe that this is the *only* fifty-year-old woman in central Vermont using a cho-

250. *The Computer and Invasion of Privacy: Hearings Before a Subcomm. of the Comm. on Gov't Operations*, 89th Cong. 3 (July 26–28, 1966).

251. The congressional hearings in the late 1960s led to the passage of the Privacy Act of 1974. *The Privacy Act of 1974*, EPIC, http://epic.org/privacy/1974act (last visited Dec. 21, 2011) [hereinafter *EPIC Privacy Act Report*]. This law bars government agencies from collecting, sharing, and retaining information that is not necessary for carrying out official duties. 5 U.S.C. § 552a(b) (2006). But the Privacy Act is the result of an odd collection of compromises, *EPIC Privacy Act Report*, *supra*, so its ability to protect against the creation of data profiles is limited. It contains a number of exceptions, including the routine use exemption (which is arguably the exception that swallows the rule), *id.* § 552a(b)(3), and exceptions for law enforcement investigations, *id.* § 552a(b)(7). For a criticism of the routine use exemption, see Paul M. Schwartz, *Privacy and Participation: Personal Information and Public Sector Regulation in the United States*, 80 IOWA L. REV. 553, 584–87 (1995), and Robert Gellman, *Does Privacy Law Work?*, *in* TECHNOLOGY AND PRIVACY: THE NEW LANDSCAPE 193, 198 (Philip E. Agre & Marc Rotenberg eds., 1997).

252. *See supra* Part III.

253. This example comes from a dissenting opinion from a recent medical privacy lawsuit. *See* IMS Health Inc. v. Sorrell, 630 F.3d 263, 283 (2d Cir. 2010) (Livingston, J., dissenting), *aff'd*, 131 S. Ct. 2653 (2011).

lesterol-lowering drug. The Internet provides a lot of information about a lot of people, but it is not a source of comprehensive and systematic information, so it is a flawed tool for the malicious intruder. At best, the adversary might be able to use some statistical source of medical treatment rates to estimate the likelihood that the Montpelier woman is unique.

Ohm and other critics of anonymization believe that once adversaries are able to sync up one anonymized database to identities, they will be able to match the combined database to a third anonymous database, and then a fourth, et cetera, until a complete profile is built.[254] This threat is premised on perfect matching attacks that contain no false matching error. If a re-identification attack is assumed to have error (which it most certainly will in the absence of a complete population registry of some sort), then the quality of the dossier will be so poor as to undermine its threat. Even in the unlikely scenario where each re-identification attack contains only a ten percent false match rate, twenty-seven percent of the observations in the combined dataset will likely contain errors.[255]

Even ignoring the snowballing error rates, the value to an adversary of anonymized data erodes over time. If adversaries are able and willing to make entropic re-identification attacks in the future, anonymized data from today will have vanishing value as time trots on for two reasons. First, people's attributes change, so making matches will be increasingly hard and subject to false positives and false negatives. Studies on databases that are known to cover the same population are, in fact, frequently difficult to match up because the subject's contemporaneous responses to the same or similar questions are often incompatible.[256] And since the profiles used to make the match will likely be riddled with error, matching to old data will often fail.[257] Second, even if a successful match is made and is verifiable, there will be less intrinsic value to knowing old attributes. No matter what the adversary's bad motives are, the value of old data (again, its marginal utility) decreases with time.

---

254. *See* Ohm, *supra* note 4, at 1725–27. Likewise, EPIC has the same conviction, claiming that the harms caused by the release of the (non-)anonymized AOL search query data will increase over time since re-identifying more AOL subjects will be easier as more and more data enters the public domain. *See Re-identification*, *supra* note 5.

255. $(0.9)^3 = 0.73$. And, of course, the adversary will not know which twenty-seven percent of entries contain the expected errors.

256. Müller, et al., *supra* note 127.

257. Even commercial data aggregation, which has the luxury of linking *identified* information, is riddled with error. Joel Stein documented the false information in his own commercial profiles in a recent *Time* article. Joel Stein, *Your Data, Yourself*, TIME, Mar. 21, 2011, at 40. Though the profiles are useful for advertising purposes, they suggest that a "database of ruin" is a fantasy well out of reach. One of the commercial databases believed that Joel Stein was an eighteen- to nineteen-year-old woman. *Id.*

Privacy advocates tend to take on the role of doom prophets — their predictions of troubles are ahead of their time.[258] Convinced of the inevitability of the harms, privacy scholars are dissatisfied with reactive or adaptive regulation and insist on taking prospective, preemptive action.[259] Dull as it is, reactive legislation is the most appropriate course for anonymized research data. Legislation inhibiting the dissemination of research data would have guaranteed drawbacks today for the research community and for society at large. We should find out whether re-identification risk materializes before taking such drastic measures.

E. Improving the Status Quo

In the meantime, we would do well to clean up the muddled state of the PII-based privacy system currently in place. Right now case law and regulatory guidance are so reluctant to commit to a protocol that data producers cannot be sure what is expected of them.

The regulatory goal of a PII-based privacy statute is quite straightforward: a data user should not be able to learn something new about a data subject using publicly available auxiliary information. Direct identifiers are removed, of course, and some additional precautions are often required. The mandates of current privacy statutes can be met using what I will refer to as the "Four Key Principles" of PII-based anonymization. These principles are not beyond the capabilities of a FOIA officer at a public agency:

(1) *Unknown Sampling Frame* — If the data producer is confident that data users cannot use public information to determine whether somebody is in the dataset or not, the other precautions described in this section need not be taken.[260]

(2) *Minimum Subgroup Count* — This concept is incorporated into my proposal above: the data producer ensures that no combination of indirect identifiers yields fewer than a certain threshold number of observations. The data producer must use good judgment in categorizing the variables as indirect identifiers or non-identifiers.[261]

---

258. Occasionally this kind of prediction is accomplished by reminiscing about simpler times. Jeffrey Rosen, for example, believes the Internet compares unfavorably to the villages described in the Babylonian Talmud. Jeffrey Rosen, *The End of Forgetting*, N.Y. TIMES, July 25, 2010, § MM (Magazine), at 30.

259. William McGeveran was quoted as making this critique in a recent New York Times article. Natasha Singer, *Technology Outpaces Privacy (Yet Again)*, N.Y. TIMES, Dec. 11, 2010, at BU3.

260. See *supra* text accompanying notes 229–234 for a description of sampling frames and how they can be used to strengthen the anonymization of data.

261. See *supra* text accompanying notes 226–228. The toughest choices will involve information that is frequently the subject of self-revelation on the Internet (e.g., preferences or movie ratings). *Id.* Also, replacing indirect identifiers with random codes does not automatically convert an indirect identifier into a non-identifier. *See supra* note 73.

(3) *Extremity-redacting* — Data producers can redact the highest or lowest value of sensitive continuous variables (e.g., income or test scores) within each subgroup if they are concerned that an adversary would be able to draw conclusions about the maximum (or minimum) value for a whole subgroup. To understand the risk this approach averts, suppose a school wishes to release a dataset containing the race, gender, and grade point average ("GPA") of its students. Suppose also that all white females at the school earned GPAs lower than 3.0. An adversary could use the database to learn that a particular white female (indeed, any white female) had a GPA below 3.0. Thus, even though the adversary cannot re-identify a particular line of data, he has learned something new and sensitive about each individual white female. If the school had redacted the highest GPA within each race-gender subgroup and replaced it with a random alphanumeric symbol, the adversary no longer knows the upper bound in the white females' (or any other group's) GPAs.[262]

(4) *Monitoring Future Overlapping Data Releases* — Finally, a data producer must ensure that it will not disclose two datasets covering the same population that can be linked through non-identifiers. Building on the race, gender, and high school GPA database example in the last paragraph, suppose the same school released a second dataset providing high school GPA and ZIP code. On its own, the second dataset seems perfectly innocuous. But any observation with a unique GPA (most likely at the bottom or top of the GPA distribution) could be linked to the first database. By doing so, an adversary can learn the race, gender, ZIP code, and GPA for those observations. This greatly increases the chance of re-identification.[263]

The theoretical concepts required to create a low-risk public dataset are not difficult when they are explained clearly and deliberately. But to this point, the judiciary has had great difficulty reasoning through and applying anonymization concepts in a principled, replicable way. The case law often contradicts itself and establishes ad hoc rules that are under- or over-protective. Even when a case reaches the

---

262. Top-coding is frequently used on income data, for a slightly different purpose than I discuss here. Income is a variable that can be used as an indirect identifier when the value is extremely high. While most people are not identifiable by their income, the very richest members of a community might be. Top-coding income to prevent this re-identification risk preserves k-anonymity and is a form of subgroup cell size control. Thus, income top-coding recodes more than just the highest income. The *Checklist on Disclosure Potential of Proposed Data Releases*, prepared by the Federal Committee on Statistical Methodology, suggests top-coding the upper limit of income distributions. Working Paper No. 22, *supra* note 109, at 103. Additional measures may be taken if a subgroup is too homogeneous with respect to a sensitive attribute.

263. Databases rarely cover the same populations since data producers have noted the high risk of overlapping disclosures on the same sample population. *See* Working Paper No. 22, *supra* note 109, at 82.

correct outcome, the analysis is often incomplete or inarticulate in its reasoning.

Consider the opinion from *Fish v. Dallas Independent School District*, discussed at length in Part II. Though the opinion applies the PII framework and properly finds that the requested dataset would run afoul of FERPA, the opinion uses flawed reasoning. The court focuses on the fact that one expert witness was able to use publicly available information to trace the identities of 550 of the Dallas students at one of the elementary schools in "less than one minute."[264] Processing speeds bear no relation to the relative ease or difficulty of re-identifying a person in a dataset. It is the discretionary decision making that comes before the computation — the skill and special information (if any) known by the human writing the attack code — that determines whether a dataset is at risk of re-identification or not.

Other cases do worse by mechanically applying statistical rules in inappropriate circumstances.[265] Consider the case of *Long v. IRS*.[266] At the trial level, the plaintiff succeeded in enforcing an old consent decree that required the Internal Revenue Service ("IRS") to release statistical reports to the plaintiff and to the public at large.[267] The issue in the case was whether one particular table that reported the number of hours spent auditing tax returns and the additional tax dollars collected through those audits violated the privacy rights of the audited taxpayers.[268] The statistics were broken down according to type of tax return, industry, and the income level of the audited taxpayer.[269]

The IRS argued that the table violates taxpayer privacy because the table contained "cells of one" — cells that described a single audited taxpayer.[270] In other words, the IRS argued that the table would violate the principle of minimum subgroup size. The plaintiff countered by arguing that "a reader would not be able to identify the taxpayer unless he already knew that the taxpayer had been audited in the relevant time period."[271] That is to say, the plaintiff was arguing that the table had an unknown sampling frame so that, in the absence of special information, an adversary would not know who was audited,

---

264. Fish v. Dallas Indep. Sch. Dist., 170 S.W.3d 226, 231 (Tex. App. 2005).

265. The California Supreme Court recently came to the preposterous holding that ZIP codes, alone, constitute "personal identification information." Pineda v. Williams-Sonoma Stores, Inc., 246 P.3d 612, 615, 618 (Cal. 2011). The defendant had used ZIP codes in conjunction with names in order to find the addresses of customers (and then used the data for marketing purposes). *Id.* at 615. The court could have solved this consumer privacy problem by ruling that ZIP codes, *when combined with names*, constituted PII. Instead they expanded the definition of PII to absurd proportions by finding that ZIP codes alone are PII. *Id.* at 620.

266. 395 F.App'x 472 (9th Cir. 2010).

267. Long v. IRS, No. C74-724P, 2006 WL 1041818, at *6 (W.D. Wash. Apr. 3, 2006).

268. *Id.* at *3.

269. *Id.*

270. *Id.*

271. *Id.*

and thus could not know who was being described in the table. So, even if the table reported the audit outcome for just one medical doctor, an adversary would not be able to determine which of the country's many medical doctors had been audited. The IRS responded that publicly available information, such as press releases or public Securities and Exchange Commission ("SEC") filings, could be used to determine the identities of some taxpayers in the sampling frame.[272] The trial court found that the IRS's position was "speculative at best," and noted that the government had provided no evidence to support its claim that a cell of one could be combined with public information to identify a taxpayer.[273] The district court properly focused on whether the sampling frame was sufficiently unknown and made a factual determination in the plaintiff's favor.

The Ninth Circuit reversed and left an illogical and unsound precedent in its wake. First, the appellate court mischaracterized the district court's opinion, claiming that the lower court had considered the table to be effectively anonymized once direct identifiers had been removed.[274] Having constructed this straw man, the appellate court went too far in knocking it down: "[W]e hold that tax data that starts out as confidential return information associated with a particular taxpayer maintains that status when it appears unaltered in a tabulation with only the identifying information removed."[275] The court determined that cells of two, on the other hand, do not implicate privacy concerns.[276] The Ninth Circuit has created a test (no cells of one) that will be over- and under-inclusive in targeting re-identification risk. The court applies a threshold that is too low for minimum subgroup size (two, as compared to the standard thresholds over three) without any regard for the protective power of the unknown sampling frame.

The unknown sampling frame principle is at the root of much confusion in U.S. privacy policy.[277] Government agencies assigned with the task of providing guidance to data producers have bungled their efforts in this regard. For example, in discussing cell size limitations, Working Paper No. 22 — a guideline for federal data disclosures — provides the following as an illustration of an aggregated statistical table with disclosure risk:

---

272. *Id.*

273. *Id.* at *4.

274. Long v. IRS, 395 F.App'x 472, 475 (9th Cir. 2010).

275. *Id.*

276. *Id.* at 475–76.

277. Some courts have gotten it right. *See, e.g.*, Conn. Dep't of Admin. Servs. v. Freedom of Info. Comm'n, No. CV 95550049, 1996 WL 88490 (Conn. Super. Ct. Feb. 9, 1996) (finding that a table showing the percentage of job applicants for a librarian position that identified themselves as having a physical handicap was not privacy-violating because the pool of applicants could not be identified).

Table 1: Number of Delinquent Children by County and Education Level of Household Head[278]

Education Level of Household Head					
County	Low	Medium	High	Very High	Total
Alpha	15	**1***	**3***	**1***	20
Beta	20	10	10	15	55
Gamma	**3***	10	10	**2***	25
Delta	12	14	7	**2***	35
Total	50	35	30	20	135

The highlighted cells are supposedly problematic, because they contain fewer than five respondents.[279] But the Federal Committee on Statistical Methodology ("FCSM") mindlessly applied the minimum subgroup count rule without grounding it in a principled theory. It is true that only one of the delinquent children lives in Alpha County with a medium-educated head of household. But that delinquent child is not in danger of being re-identified. An adversary has no way of knowing who is in this sample of delinquent children unless the adversary already knows the child is delinquent. Knowing that some child lives in Alpha County with a medium-educated head of household also tells the adversary nothing about whether that child is delinquent because he cannot determine whether that child is in the sample. If the adversary did know that some particular target is in the sample, he would already know the most potentially harmful information about the target: that the target is a delinquent child.[280] When the conditions of an unknown sampling frame are met, the cell sizes have no relation to the hypothetical abuses that could flow from tabular data.[281]

In another brief, FCSM suggests that the problem with small cells in a simple frequency table like this is that anyone privy to the infor-

278. Working Paper No. 22, *supra* note 109, at 16.

279. *Id.*

280. The table could pose problems if there are very few highly educated parents in a given county. Suppose, for example, that Alpha County had only one head of household with very high education. Then members of the community might be able to discern that the head of household in question has a delinquent child. The definition of "unknown sampling frame" provided earlier in this section guards against these scenarios.

281. In the discussion of *Southern Illinoisan* in Part III, I discuss how an aggregated table can be used to slightly increase the chance of re-identification when used by a sophisticated adversary (of dubitable existence), but small cell sizes are no more vulnerable than large ones for these tactics.

mation about one of the data subjects is more likely to be able to identify the other people described in the same, small cell.[282] This suggestion may sound reasonable, but it does not logically follow. Consider the parents of a delinquent child, who know without ambiguity where their child falls in the frequency table. Even if their child was one of the two delinquent children from Gamma County with a head of household with very high education, that parent could not learn anything about the identity of the other delinquent child unless they already knew the county and education level associated with that delinquent child (in which case, they would know all there is to know).

My criticism of this exemplar table is not meant to imply that tables of aggregated information cannot breach privacy. They can and they have. The following table reports pass rates for the No Child Left Behind Exit Exam for a single high school in California. This table shows how the results were reported in public documents by the California Department of Education.

Table 2: California High School Exit Exam (CAHSEE) Results for Mathematics and English Language Arts (ELA) by Gender and Ethnic Designation, (Combined 2008) for (Grade 11)

[Name of School Redacted][283]

	MATH		**ELA**	
	Took	**Passed**	**Took**	**Passed**
All Students	**27**	**3**	**23**	**3**
Female	4	n/a	3	n/a
Male	23	3	20	2
Hispanic or Latino	20	3	18	3
White	7	n/a	5	n/a

282. CONFIDENTIALITY AND DATA ACCESS COMM. & FED. COMM. ON STATISTICAL METHODOLOGY, CONFIDENTIALITY AND DATA ACCESS ISSUES AMONG FEDERAL AGENCIES 4 (2001), *available at* http://fcsm.gov/committees/cdac/brochur10.pdf ("For example, a two-dimensional frequency count table may have rows corresponding to employment sectors (industry, academia, nonprofit, government, military) and columns corresponding to income categories (in increments of $10,000). . . . Using this example, such a tabulation could result in a disclosure of confidential information if . . . only 2 cases of any sector fell into the same income category (permitting the conclusion on the part of anyone privy to the information about one of the cases, to know the income of the other).").

283. Muralidhar & Sarathy, *supra* note 168, at 9 (table reformatted by author).

This table violates privacy by revealing the math test results with certainty for female and white students, despite the school district's effort to redact results for cells smaller than ten by replacing the number with "n/a."[284]

Blame for deficient anonymization does not reside with the data-producing agencies alone. Regulators charged with the task of setting out standards for data sharing seem to go out of their way to avoid clarity.[285] Working Paper No. 22 runs through a menu of options for data producers, including random sampling, top-coding, adding random noise, and blurring or clustering the indirect identifier variables.[286] But the paper does not provide a uniform guideline, admitting that "there are no accepted measures of disclosure risk for a microdata file, so there is no 'standard' that can be applied to assure that protection is adequate."[287]

This guidance is stunningly inadequate for a small firm or public agency charged with the task of producing a public-use dataset. It is understandable that statistical agencies would not want to commit themselves to a list of indirect identifiers or to a specific fixed set of protocols. Identifying which variables are indirect identifiers requires some working knowledge of the dataset and the publicly available resources that can be matched to the dataset. But the privacy regulators fail even to elucidate workable principles.[288] The regulatory body that administers HIPAA, for example, has failed to provide clear guidance on "specific conditions that must be met in order for privacy

---

284. *Id.*

285. This is how the FPCO responded to requests for better guidance on the application of education privacy law to de-identified data:

> In response to requests for guidance on what specific steps and methods should be used to de-identify information . . . it is not possible to prescribe or identify a single method to minimize the risk of disclosing personally identifiable information in redacted records or statistical information that will apply in every circumstance . . . . This is because determining whether a particular set of methods for de-identifying data and limiting disclosure risk is adequate cannot be made without examining the underlying data sets, other data that have been released, publicly available directories, and other data that are linked or linkable to the information in question.

Family Educational Rights and Privacy, 73 Fed. Reg. 74,806, 74,835 (Dec. 9, 2008). The FPCO is abandoning its responsibility to provide guidance on anonymization practices because it cannot provide a fool-proof step-by-step instruction manual applicable to every scenario.

286. Working Paper No. 22, *supra* note 109, at 24–33.

287. *Id.* at 24.

288. The Checklist on Disclosure Potential of Proposed Data Releases succeeds in providing some guidance on the sort of issues that must be considered when preparing a public-use microdata file. *See* INTERAGENCY CONFIDENTIALITY AND DATA ACCESS GROUP, FED. COMM. ON STATISTICAL METHODOLOGY, CHECKLIST ON DISCLOSURE POTENTIAL OF PROPOSED DATA RELEASES 6–17 (1999), *available at* http://fcsm.gov/committees/cdac/checklist_799.doc. But the guidance goes over the heads of the average government administrator, unfamiliar with "sampling frame[s]," "matching," and "nesting variables." *Id.* Like the other resources, the Checklist increases concern without providing clear principles.

risks to be minim[ized]," leaving the details to be sorted out by individual privacy boards and Institutional Review Boards.[289]

The result is complete chaos. Simply, there are no standard privacy practices. Richard Sander, a law professor at University of California, Los Angeles, recently requested anonymized admissions data from 100 public colleges and 70 public law schools.[290] The requests were submitted pursuant to an effort to conduct a systematic examination of admissions practices, but the data collection process serves as its own meta-experiment on public records compliance. Since Sander sent identical requests to every school, their responses provide a unique opportunity to observe the variance in interpretations of education privacy laws. The meta-experiment produced two important insights. First, the schools had widely divergent interpretations of their obligations under FERPA. Some of the schools complied with the FOIA requests right away and without redactions, but the majority provided data only after protracted negotiations lasting as long as two years. One fifth of the schools refused even dramatically scaled-back requests that presented no appreciable risk of re-identification. Second, the diversity among state FOIA statutes and privacy laws had little bearing on a school's likelihood to provide data. Noncompliant schools shared their state borders with compliant ones. Some of the refusing schools sent letters denying the request on the basis of privacy exemptions to the state's public records laws. Other schools became nonresponsive in the course of negotiations.[291] And a few schools effectively denied the request by sending data that redacted race information or by charging excessive fees.[292]

The void in standard practices naturally heightens the fears of members of the public, who view inconsistency as evidence that their

---

289. Barbara J. Evans, *Congress' New Infrastructural Model of Medical Privacy*, 84 NOTRE DAME L. REV. 585, 626 (2009).

290. Professor Sander's raw data and other study materials are on file with the author and are being used with permission from Professor Sander.

291. These schools received several inquiries in a variety of formats, including, at the very least, two mailed letters, two e-mails, and two phone calls. The project's logs for schools that were unresponsive read like parodies of bureaucratic inefficiency. Here is an example (names and contact information redacted):

> [10/17] VW said she never got [the request], and to speak to ES. [Phone number]. Spoke to ES, told her we would resend request. ASW 6/5: Letter mailed and emailed to ES. ASW 6/13: Recd email from ES acknowledging request and advising that it would be more than $150; they will advise us of the cost soon. TP 8/15/8 spoke with ES who said she does not remember our request but will check on it and get back to us. TP send a follow-up e-mail to [email address] // 11/19/08 ES assistant said she is out of the office for the week. Lft msg on voice mail.//12/09/08 TP got a hold of ES who connected me to vice Chancelor JP. JP asked that I e-mail him the requests. I did on same day.//1/9/9 TP left phone message for JP.

292. The University of Maryland Law School invoiced Professor Sander $3,700 for the data — an amount thirty-seven times the average cost estimates.

confidentiality may not be sufficiently protected.[293] The discrediting of anonymization and the growing perception that current privacy protocols are a fragile facade have already taken a toll on the data commons. Some public-use datasets require researchers to sign notarized affidavits and cut through a good deal of red tape before and during their use of the data.[294] And some agencies have pulled public datasets into on-site research enclaves.[295] These trends increase the costs of doing research. Some policymakers are interfering with agencies' ability to release research data at all: the Department of Transportation and Related Agencies Appropriations Act was the first federal law prohibiting access to records in the absence of individual opt-in consent, even though the records were previously open to the public and had not been the subject of any known abuses.[296] Conditioning the collection of certain categories of information on the consent of the consumer is fatal to the collection of any reasonably useful data.[297]

The stakes for data privacy have reached a new high-water mark, but the consequences are not what they seem. We are at great risk not of privacy threats, but of information obstruction.

VI. CONCLUSION: THE TRAGEDY OF THE DATA COMMONS

The contours of the right to privacy are in the grips of an existential crisis. Social networking, history-sniffing cookies, and costless

---

293. CHARLES J. SYKES, THE END OF PRIVACY 135 (1999) (noting that, in the context of medical privacy, "[i]n an age where . . . medical datawebs cover the country from coast to coast, only uniform standards have any reasonable prospect of assuring patient confidentiality"); Andrew B. Serwin, *Privacy 3.0 — The Principle of Proportionality*, 42 U. MICH. J.L. REFORM 869, 875 (2009) (finding that inconsistent legal standards cannot meet society's need for privacy).

294. The instruction manual for applying to use a dataset held by the National Center for Education Statistics is fifty-six pages long. *See* INST. OF EDUC. SCIS., U.S. DEP'T OF EDUC., RESTRICTED-USE DATA PROCEDURES MANUAL (2011), *available at* http://nces.ed.gov/pubs96/96860rev.pdf.

295. The National Center for Health Statistics ("NCHS") changed their data access policies in 2005 and pulled some previously public data files into a research enclave that requires pre-approval and the payment of a fee. *See NCHS Data Release and Access Policy for Micro-data and Compressed Vital Statistics Files*, CENTERS FOR DISEASE CONTROL AND PREVENTION, http://www.cdc.gov/nchs/nvss/dvs_data_release.htm (last updated Apr. 26, 2011). For a description of the process to apply for access to the research enclave, see *NCHS Research Data Center*, CENTERS FOR DISEASE CONTROL AND PREVENTION, http://www.cdc.gov/rdc (last updated Nov. 3, 2009).

296. Department of Transportation and Related Agencies Appropriations Act, Pub. L. No. 106-69, § 350, 113 Stat. 986, 1025–26 (1999); Cate, *supra* note 9, at 12.

297. Cate, *supra* note 9, at 15. Opt-in requirements produce insurmountable selection bias problems because the people who opt into the study (or those that do not) often share characteristics. Researchers cannot assume that the subjects who have chosen to opt in are typical or representative of the general population. Bas Jacobs, Joop Hartog & Wim Vijverberg, *Self-Selection Bias in Estimated Wage Premiums for Earnings Risk*, 37 EMPIRICAL ECON. 271, 272 (2009).

digital archiving have forced us to grapple with new and difficult problems. There are many worthy targets for the worries of privacy scholars. Research data is not one of them.

Parts II–IV of this Article analyzed the risk and the utility of public research data. With high benefit and low risk, the inescapable conclusion is that current privacy risks have little to do with anonymized research data, and the sharing of such data should be aided by the law rather than discouraged by it. But the proposals in Part V will no doubt be controversial. Now that researchers, legal scholars, and major policymakers have converged on an alarmist interpretation of the current state of data sharing, cool-headed balancing between risks and benefits is extraordinarily difficult. Our collective focus has been set on detriment alone.

Paul Ohm refers to the "inchoate harm[s]" of datasets that are released without airtight protections against re-identification.[298] Conceived of this way, the right to not be re-identified is one that need not bend to *any* considerations for the public interest in reliable research data. Ohm's approach to privacy policy is the same as my own — he advocates a balancing of the interests in privacy against the interests in data release.[299] Ohm and I arrive at very different policy proposals because we have divergent estimations of re-identification risks and the value of public data releases. However, other scholars have encouraged privacy law to drift into a property-based enforcement regime.[300] Proponents of property entitlement would say, "It is *my* data, and I want it out of the data commons." To conclude this Article, I highlight the features that make a property regime in anonymized data unworkable and unwise. Because risk is borne by individuals while utility is spread across the entire community, circumstances are ripe for a tragedy of the commons. The tort liability model for enforcement of privacy rights is much more sensible since tort liability rules are tailored to the risks and costs at a higher level of generality — the societal level.

A. Problems with the Property Model

There is no Pareto-optimal way to share data. This, unfortunately, is irrefutable. Though we are collectively better off with public re-

298. Ohm, *supra* note 4, at 1749. The term "inchoate harm" is inappropriate in the context of research data. It evokes images of a loaded gun — something nefarious and unnecessarily dangerous. Privacy harms can be described as "inchoate" when a sensitive piece of information has been exposed to public view, and it is unclear whether or when it will be harmfully linked to a data subject. *See id.* at 1749–50. This is an excellent approach for data spills (the accidental release of identifiable data). But in anonymized form, research data is no different from the data banks sitting on a server or even a personal computer. While it is susceptible to an intervening wrong, its existence is not, in itself, wrongful.

299. Ohm, *supra* note 4, at 1736.

300. *See supra* note 6.

search data, sharing data imposes risk on the data subjects. This risk can be greatly reduced by taking certain precautions, but it can never reach zero. Who, then, is to decide how much risk is too much?

Many people want (and probably believe they have) a property interest in information that describes them.[301] The practical significance of enforcing privacy rights through the property model is that the data subject retains the right to hold out. Thus, recent class action lawsuits for releasing research data demanded injunctions against sharing data in the future and brought claims for trespass to chattels.[302] Additionally, Lawrence Lessig, Jerry Kang, and Paul Schwartz have argued that Americans should have control over their information that is at least as strong as a property regime would permit and preferably stronger.[303]

In the case of research data, the property model is the wrong choice, not only for efficiency reasons, but also because it fails to meet the distributional goals required for justice.[304] Americans are naturally distrustful about data collection. Significant segments of the population continue to evade U.S. Census reporting, despite both the legal mandate to do so[305] and the Bureau's clean confidentiality record during the last six decades.[306] If data subjects refuse to consent to even small amounts of risk, which a rational actor model would predict they would do, then the data commons will dwindle as property is claimed.[307]

---

301. For example, in the complaint of a lawsuit against Apple for the disclosure of data (which Apple claims was anonymized), the data was described as "confidential information and personal property that [the data subjects] do not expect to be available to an unaffiliated company." Complaint at 5, Lalo v. Apple Inc., 2010 WL 5393496 (N.D. Cal. Dec. 23, 2010) (No. 5:10-cv-05878-PSG) [hereinafter Apple Complaint].

302. *See, e.g.*, In re Pharmatrak, Inc., 329 F.3d 9, 16 (1st Cir. 2003); Apple Complaint, *supra* note 301.

303. *See supra* note 6. Paul Schwartz challenges a simple property model for information privacy by noting that consumers will foreseeably sell their alienable information for too little compensation. Schwartz, *supra* note 6, at 2091. Schwartz embraces many of the aspects of a property model, but also proposes that government regulation should provide a right of exit (or claw-back) and a realm of inalienability. *Id.* at 2094–116.

304. The sound choice between liability and property rules will look to both efficiency and distributional goals. Guido Calabresi & A. Douglas Melamed, *Property Rules, Liability Rules, and Inalienability: One View of the Cathedral*, 85 HARV. L. REV. 1089, 1110 (1972) (explaining how liability rules facilitate the combination of efficiency and distributive results which would be difficult to achieve under property rules).

305. Eleanor Singer, Nancy A. Mathiowetz & Mick P. Couper, *The Impact of Privacy and Confidentiality Concerns on Survey Participation: The Case of the 1990 U.S. Census*, 57 PUB. OPINION Q. 465, 479 (1993).

306. While the U.S. Census Bureau has had no recent (known) confidentiality breaches, the Bureau did transfer confidential records to the U.S. Department of Justice during World War II to facilitate identifying and rounding up Japanese-Americans and placing them into internment camps. *See* JR Minkel, *Confirmed: The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II*, SCI. AM. (Mar. 30, 2007), http://www.scientificamerican.com/article.cfm?id=confirmed-the-us-census-b.

307. *See* Hardin, *supra* note 8, at 1244. Each data subject will view their decision to take their own data out of the commons as the optimal choice: the data commons is rich enough

This problem is analogous to the modern vaccine controversy. Children under the age of vaccination are often at the greatest risk of death from virulent diseases like whooping cough.[308] The best protection is for everyone else (of eligible age) to get the vaccine, even though the vaccine itself poses dubious but popularly accepted risks.[309] Parents who choose not to vaccinate their children expect to have it both ways: since everyone else is vaccinated, their child is unlikely to be exposed to the virus or disease. But they also avoid the small chance that their child could have an adverse reaction to the vaccination. The trouble is, once enough parents opt out of the vaccination pool, the communal protection falls apart. Thus, we are now witnessing a resurgence in infant mortality from whooping cough because the virus is spreading among adults and older children, who historically had been vaccinated but no longer are.[310]

Like the communal vaccination shield, the data commons is especially vulnerable to opt-outs. As people opt out, the value of the overall data diminishes precipitously rather than linearly: even a small number of holdouts will produce selection bias effects that compromise the utility of the remaining data. Khaled El Emam, Elizabeth Jonker, and Anita Fineberg have recently compiled and analyzed the evidence of selection bias caused by consent requirements to perform research on observation health data — data that was already collected in the course of treatment, such that research requires no additional interaction with the patients.[311] Consent is denied more frequently by patients who are younger, African American, unmarried, less educated, of lower socio-economic status, or — importantly — healthy.[312] These patterns are very difficult to control for, and they cause distortions in health research.[313] Put bluntly, property rights that follow the information into the data commons (and allow the data to be clawed

---

to allow for research, but their own data is not exposed to risk of re-identification. If many people arrived at this same choice in the course of their own independent evaluations, there would be no commons left. I discuss the differences between the data commons and the traditional tragedy of the commons in Part I. *See supra* note 8.

308. *See generally Pertussis (Whooping Cough)*, CENTERS FOR DISEASE CONTROL & PREVENTION, http://www.cdc.gov/pertussis/index.html (last updated Aug. 22, 2011).

309. *See* Chris Mooney, *Why Does the Vaccine/Autism Controversy Live On?*, DISCOVER MAG., June 2009, at 58, 58–59.

310. Ijeoma Ejigiri, *The Resurgence of Pertussis: Is Lack of Adult Vaccination to Blame?*, CLINICAL CORRELATIONS (Feb. 23, 2011), http://www.clinicalcorrelations.org/?p=3951.

311. Khaled El Emam et al., The Case for De-identifying Personal Health Information 21–29 (Jan. 18, 2011) (unpublished manuscript), *available at* http://ssrn.com/abstract=1744038.

312. *Id.* at 27.

313. *Id.* at 25–28.

back out) would allow holdouts to wreak disproportional havoc on research.[314]

The impulse to enforce research data privacy rights through property rules should be jettisoned and a tort approach restored.[315] On this issue, Paul Ohm and I agree that the public interest is best served by asking whether the utility of a public dataset significantly outweighs the risk of harm.[316] This would mark a return to the rational balancing anticipated by Samuel Warren and Louis Brandeis, who recognized that privacy rights should not interfere with information flow when that information is socially valuable.[317] This balancing of risks and benefits will also realign the policy discourse with the anonymization practices that are already widely in use and embraced by privacy experts in the statistics and social science fields. Anonymization was never believed to be a "privacy-providing panacea."[318] As Douglas Sylvester and Sharon Lohr correctly assert, "[t]he law, in fact, does not require that there be absolutely no risk that an individual could be identified from released data."[319] Rather, the law was assumed to reflect a conservative position in the risk-utility analysis — and it still does.

Radical as they may sound, this Article's proposals are formally reconcilable with the privacy scholarship that demands inalienable rights in the control of information. De-identified (anonymized) data need not be considered as relating to the underlying data subject at all — unless and until their data has been re-identified. The theoretical foundations for establishing a distinct regime for anonymized data are already in existence. Jerry Kang has noted that privacy is in some tension with intellectual property since there is no available copyright ownership interest in facts.[320] Once data has been unlinked from an identifiable person, perhaps it is best understood as a fact in the public domain. Better still, Ted Janger and Paul Schwartz have proposed a

---

314. If a property rule is crafted to avoid over-protection then it will likely end up in a form that is under-protective. Suppose we were to determine that the data subject had alienated his right to the information as soon as he gave it to the data producer (say, a retailer or his doctor), then a property regime would constrain the state from interfering with the data producer's use, no matter how badly the original data subject was under-compensated. This is not sound policy in the majority of contexts in which data is collected — where the information is given for a purpose without concrete attention to the additional uses (in identifiable form or not) to which the data will be put.

315. Fred Cate has argued that democratic values would benefit from a shift away from property rights, though he sees value, often overlooked by the legal academy, in allowing private entities to use data for secondary uses. *See* Cate, *supra* note 9, at 12.

316. Ohm advises regulators to compare the risks of unfettered information flow to its likely costs in privacy. Ohm, *supra* note 4, at 1768.

317. Warren & Brandeis, *supra* note 43, at 214.

318. Ohm, *supra* note 4, at 1716.

319. Douglas J. Sylvester & Sharon Lohr, *Counting on Confidentiality: Legal and Statistical Approaches to Federal Privacy Law after the USA Patriot Act*, 2005 WIS. L. REV. 1033, 1113 (2005).

320. Kang & Buchner, *supra* note 6, at 233.

move to "constitutive privacy" rights, where access to information and limits on it should be modeled with an eye toward the nature of our society and the way we like to live.[321] Here the "democratic community"[322] is much better served by relinquishing an individual's control over anonymized research data.

Detaching privacy rights from anonymized data presents the best option available because it prevents what Anita Allen calls the maldistribution of privacy.[323] Consider the following scenario: a school district wishes to test a theory that implicit biases cause its teachers to depress grades of minority students when students are evaluated on subjective criteria. To test the hypothesis, the school district uses the objective scores received by its students on validated exams as controls to see if minority students receive significantly lower grades when grading is left to the teacher's subjective judgment. A small set of parents, after catching wind of the study, object to the use of their (Caucasian) children's data because the secondary use of their children's information does not suit their interests. Should we consider the data, in anonymized form, to be *their* data? Individuals' control over research data would result in a maldistribution of knowledge.

B. The Data Subject as the Honorable Public Servant

The data commons is the tax we pay to our public information reserves. Danielle Citron and Paul Schwartz have persuasively argued that privacy is a critical ingredient to a healthy social discourse.[324] In many respects this is true, but if taken to the extreme, data privacy can also make discourse anemic and shallow by removing from it relevant and readily attainable facts.

In time, technological solutions are likely to pare down the existing tension between data utility and disclosure risk.[325] Statistical software that allows the dataset to remain on a secure server while researchers submit statistical queries has been developed, and many

---

321. Edward J. Janger & Paul M. Schwartz, *The Gramm-Leach-Bliley Act, Information Privacy, and the Limits of Default Rules*, 86 MINN. L. REV. 1219, 1250–51 (2002).

322. *Id.* at 1251.

323. Anita L. Allen, *Coercing Privacy*, 40 WM. & MARY L. REV. 723, 725 (1999) ("Neither privacy nor private choice, however, is an absolute, unqualified good. There can be too much privacy, and it can be maldistributed.").

324. Danielle Keats Citron, *Fulfilling Government 2.0's Promise with Robust Privacy Protections*, 78 GEO. WASH. L. REV. 822, 841–43 (2010); Schwartz, *supra* note 251, at 593 (1995) ("The boundless collection, processing, and dissemination of personal data can have a deleterious effect on the ability of individuals to join in social discourse.").

325. John M. Abowd & Julia Lane, *New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers* 3050 PRIVACY IN STATISTICAL DATABASES: LECTURE NOTES IN COMPUTER SCIENCE 282, 283 (2004), *available at* http://www.springerlink.com/content/27nud7qx09qurg3p/fulltext.pdf.

data producers are slowly beginning to implement it.[326] In the meantime, anonymization continues to be an excellent compromise. Rather than sounding alarms and feeding into preexisting paranoia, the voices of reason from the legal academy should invoke a civic duty to participate in the public data commons and to proudly contribute to the digital fields that describe none of us and all of us at the same time.

326. For example, the U.S. Census Bureau's American FactFinder service allows users to submit queries for the creation of customized tables. *American FactFinder*, U.S. CENSUS BUREAU, http://factfinder.census.gov/servlet/DatasetMainPageServlet (last visited Dec. 21, 2011).