BROKEN PROMISES OF PRIVACY: RESPONDING TO THE SURPRISING FAILURE OF ANONYMIZATION

Paul Ohm*

Computer scientists have recently undermined our faith in the privacy-protecting power of anonymization, the name for techniques that protect the privacy of individuals in large databases by deleting information like names and social security numbers. These scientists have demonstrated that they can often "reidentify" or "deanonymize" individuals hidden in anonymized data with astonishing ease. By understanding this research, we realize we have made a mistake, labored beneath a fundamental misunderstanding, which has assured us much less privacy than we have assumed. This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention. We must respond to the surprising failure of anonymization, and this Article provides the tools to do so.

---

* Associate Professor, University of Colorado Law School. This Article was presented at the Privacy Law Scholars Conference and at conferences and faculty workshops at Harvard's Center for Research and Computer Science and Berkman Center, Princeton's Center for Information Technology Policy, Fordham University Center for Law and Information Policy, University of Washington School of Law, University of Washington's Computer Science & Engineering Department, NYU Information Law Institute, DePaul Center for IP Law and Information Technology, International Association of Privacy Professionals Global Privacy Summit, and the University of Colorado Law School. I thank all participants for their comments.

Thanks in particular to Caspar Bowden, Ramon Caceres, Ryan Calo, Deborah Cantrell, Danielle Citron, Nestor Davidson, Pierre de Vries, Vasant Dhar, Cynthia Dwork, Jed Ela, Ed Felten, Victor Fleischer, Susan Freiwald, Brett Frischmann, Michael Froomkin, Simson Garfinkel, Lauren Gelman, Eric Goldman, James Grimmelmann, Mike Hintze, Chris Hoofnagle, Clare Huntington, Jeff Jonas, Jerry Kang, Nancy Kim, Jon Kleinberg, Sarah Krakoff, Tim Lee, William McGeveran, Deven McGraw, Viva Moffat, Tyler Moore, Arvind Narayanan, Helen Nissenbaum, Scott Peppett, Jules Polonetsky, Foster Provost, Joel Reidenberg, Ira Rubinstein, Andrew Schwartz, Ari Schwartz, Vitaly Shmatikov, Chris Soghoian, Dan Solove, Latanya Sweeney, Peter Swire, Salil Vadhan, Michael Waggoner, Phil Weiser, Rebecca Wright, Felix Wu, and Michael Zimmer for their comments. This research was supported by a pre-tenure research leave grant by the University of Colorado Law School, and for this I thank Dean David Getches and Associate Dean Dayna Matthew. Finally, I thank my research assistant, Jerry Green.

INTRODUCTION

Imagine a database packed with sensitive information about many people. Perhaps this database helps a hospital track its patients, a school its students, or a bank its customers. Now imagine that the office that maintains this database needs to place it in long-term storage or disclose it to a third party without compromising the privacy of the people tracked. To eliminate the privacy risk, the office will *anonymize* the data, consistent with contemporary, ubiquitous data-handling practices.

First, it will delete personal identifiers like names and social security numbers. Second, it will modify other categories of information that act like identifiers in the particular context—the hospital will delete the names of next of kin, the school will excise student ID numbers, and the bank will obscure account numbers.

What will remain is a best-of-both-worlds compromise: Analysts will still find the data useful, but unscrupulous marketers and malevolent identity thieves will find it impossible to identify the people tracked. Anonymization will calm regulators and keep critics at bay. Society will be able to turn its collective attention to other problems because technology will have solved this one. Anonymization ensures privacy.

Unfortunately, this rosy conclusion vastly overstates the power of anonymization. Clever adversaries can often *reidentify* or *deanonymize* the people hidden in an anonymized database. This Article is the first to comprehensively incorporate an important new subspecialty of computer science, reidentification

science, into legal scholarship.[1] This research unearths a tension that shakes a foundational belief about data privacy: *Data can be either useful or perfectly anonymous but never both*.

Reidentification science disrupts the privacy policy landscape by undermining the faith we have placed in anonymization. This is no small faith, for technologists rely on it to justify sharing data indiscriminately and storing data perpetually, while promising users (and the world) that they are protecting privacy. Advances in reidentification expose these promises as too often illusory.

These advances should trigger a sea change in the law because nearly every information privacy law or regulation grants a get-out-of-jail-free card to those who anonymize their data. In the United States, federal privacy statutes carve out exceptions for those who anonymize.[2] In the European Union, the famously privacy-protective Data Protection Directive extends a similar safe harbor through the way it defines "personal data."[3] Yet reidentification science exposes the underlying promise made by these laws—that anonymization protects privacy—as an empty one, as broken as the technologists' promises. At the very least, lawmakers must reexamine every privacy law, asking whether the power of reidentification and fragility of anonymization have thwarted their original designs.

The power of reidentification also transforms the public policy debate over information privacy. Today, this debate centers almost entirely on squabbles over magical phrases like "personally identifiable information" (PII) or "personal data." Advances in reidentification expose how thoroughly these phrases miss the point. Although it is true that a malicious adversary can use PII such as a name or social security number to link data to identity, as it turns out, the adversary can do the same thing using information that nobody would classify as personally identifiable.

---

1. A few legal scholars have considered the related field of statistical database privacy. *E.g.* Douglas J. Sylvester & Sharon Lohr, *The Security of Our Secrets: A History of Privacy and Confidentiality in Law and Statistical Practice*, 83 DENV. U. L. REV. 147 (2005); 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. In addition, a few law students have discussed some of the reidentification studies discussed in this Article, but without connecting these studies to larger questions about information privacy. *See, e.g.*, Benjamin Charkow, Note, *The Control Over the De-Identification of Data*, 21 CARDOZO ARTS & ENT. L.J. 195 (2003); Christine Porter, Note, *De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information*, 5 SHIDLER J.L. COM. & TECH. 3 (2008) (discussing the AOL and Netflix stories).

2. *See infra* Part II.B.

3. Council Directive 95/46 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L281) 31 [hereinafter EU Data Protection Directive].

How many other people in the United States share your specific combination of ZIP code, birth date (including year), and sex? According to a landmark study, for 87 percent of the American population, the answer is zero; these three pieces of information uniquely identify each of them.[4] How many users of the Netflix movie rental service can be uniquely identified by when and how they rated any three of the movies they have rented? According to another important study, a person with this knowledge can identify more than 80 percent of Netflix users.[5] Prior to these studies, nobody would have classified ZIP code, birth date, sex, or movie ratings as PII. As a result, even after these studies, companies have disclosed this kind of information connected to sensitive data in supposedly anonymized databases, with absolute impunity.

These studies and others like them sound the death knell for the idea that we protect privacy when we remove PII from our databases. This idea, which has been the central focus of information privacy law for almost forty years, must now yield to something else. But to what?

In search of privacy law's new organizing principle, we can derive from reidentification science two conclusions of great importance:

First, the power of reidentification will create and amplify privacy harms. Reidentification combines datasets that were meant to be kept apart, and in doing so, gains power through accretion: Every successful reidentification, even one that reveals seemingly nonsensitive data like movie ratings, abets future reidentification. Accretive reidentification makes all of our secrets fundamentally easier to discover and reveal. Our enemies will find it easier to connect us to facts that they can use to blackmail, harass, defame, frame, or discriminate against us. Powerful reidentification will draw every one of us closer to what I call our personal "databases of ruin."[6]

Second, regulators can protect privacy in the face of easy reidentification only at great cost. Because the utility and privacy of data are intrinsically connected, no regulation can increase data privacy without also decreasing data

4. Latanya Sweeney, *Uniqueness of Simple Demographics in the U.S. Population* (Laboratory for Int'l Data Privacy, Working Paper LIDAP-WP4, 2000). For more on this study, see *infra* Part I.B.1.b. More recently, Philippe Golle revisited Dr. Sweeney's study, and recalculated the statistics based on year 2000 census data. Dr. Golle could not replicate the earlier 87 percent statistic, but he did calculate that 61 percent of the population in 1990 and 63 percent in 2000 were uniquely identified by ZIP, birth date, and sex. Philippe Golle, *Revisiting the Uniqueness of Simple Demographics in the US Population*, 5 ACM WORKSHOP ON PRIVACY IN THE ELEC. SOC'Y 77, 78 (2006).

5. Arvind Narayanan & Vitaly Shmatikov, *Robust De-Anonymization of Large Sparse Datasets*, *in* PROC. OF THE 2008 IEEE SYMP. ON SECURITY AND PRIVACY 111, 121 [hereinafter *Netflix Prize Study*]. For more on this study, see *infra* Part I.B.1.c.

6. *See infra* Part III.A.

utility. No useful database can ever be perfectly anonymous, and as the utility of data increases, the privacy decreases.

Thus, easy, cheap, powerful reidentification will cause significant harm that is difficult to avoid. Faced with these daunting new challenges, regulators must find new ways to measure the risk to privacy in different contexts. They can no longer model privacy risks as a wholly scientific, mathematical exercise, but instead must embrace new models that take messier human factors like motive and trust into account. Sometimes, they may need to resign themselves to a world with less privacy than they would like. But more often, regulators should prevent privacy harm by squeezing and reducing the flow of information in society, even though in doing so they may need to sacrifice, at least a little, important counter values like innovation, free speech, and security.

The Article proceeds in four Parts. Part I describes the dominant role anonymization plays in contemporary data privacy practices and debates. It surveys the recent, startling advances in reidentification science, telling stories of how sophisticated data handlers—America Online, the state of Massachusetts, and Netflix—suffered spectacular, surprising, and embarrassing failures of anonymization. It then looks closely at the science of reidentification, borrowing heavily from a computer science literature heretofore untapped by legal scholars. Part II reveals how these powerful advances in reidentification thwart the aims of nearly every privacy law and regulation. Part III considers three simple and appealing responses to these imbalances, but ultimately rejects them as insufficient and incomplete. Finally, Part IV offers a way forward, proposing a test for deciding when to impose new privacy restrictions on information flow and demonstrating the test with examples from health and internet privacy.

I. ANONYMIZATION AND REIDENTIFICATION

A. The Past: Robust Anonymization

Something important has changed. For decades, technologists have believed that they could robustly protect people's privacy by making small changes to their data, using techniques surveyed below. I call this the *robust anonymization assumption*. Embracing this assumption, regulators and technologists have promised privacy to users, and in turn, privacy is what users have come to expect. Today, anonymization is ubiquitous.

But in the past fifteen years, computer scientists have established what I call the *easy reidentification result*, which proves that the robust anonymization

assumption is deeply flawed—not fundamentally incorrect, but deeply flawed. By undermining the robust anonymization assumption, easy reidentification will topple the edifices of promise and expectation we have built upon anonymization. The easy reidentification result will also wreak havoc on our legal systems because our faith in robust anonymization has thoroughly infiltrated our privacy laws and regulations, as Part II explores. But before we deploy the wrecking balls, this Part reviews the story of how we built these grand structures, to explain what we are about to lose.

1. Ubiquitous Anonymization

Anonymization plays a central role in modern data handling, forming the core of standard procedures for storing or disclosing personal information. What is anonymization, why do people do it, and how widespread is it?

a. The Anonymization/Reidentification Model

Let us begin with terminology. A person or entity, the data administrator, possesses information about individuals, known as data subjects. The data administrator most often stores the information in an electronic database, but it may also maintain information in other formats, such as traditional paper records.

Data administrators try to protect the privacy of data subjects by anonymizing data. Although I will later argue against using this term,[7] I am not quite ready to let it go, so for now, anonymization is a process by which information in a database is manipulated to make it difficult to identify data subjects.

Database experts have developed scores of different anonymization techniques, which vary in their cost, complexity, ease of use, and robustness. For starters, consider a very common technique: suppression.[8] A data administrator suppresses data by deleting or omitting it entirely. For example, a hospital data administrator tracking prescriptions will suppress the names of patients before sharing data in order to anonymize it.

The reverse of anonymization is reidentification or deanonymization.[9] A person, known in the scientific literature as an adversary,[10] reidentifies

7. *See infra* Part II.C.2.

8. *See* Latanya Sweeney, *Achieving* k*-Anonymity Privacy Protection Using Generalization and Suppression*, 10 INT'L J. ON UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYS. 571, 572 (2002).

9. *E.g.*, *Netflix Prize Study*, *supra* note 5, at 111–12.

10. *Id.*

anonymized data by linking anonymized records to outside information, hoping to discover the true identity of the data subjects.

b. The Reasons to Anonymize

Data administrators anonymize to protect the privacy of data subjects when storing or disclosing data. They disclose data to three groups. First, they release data to third parties: For example, health researchers share patient data with other health researchers,[11] websites sell transaction data to advertisers,[12] and phone companies can be compelled to disclose call logs to law enforcement officials.[13] Second, administrators sometimes release anonymized data to the public.[14] Increasingly, administrators do this to engage in what is called crowdsourcing—attempting to harness large groups of volunteer users who can analyze data more efficiently and thoroughly than smaller groups of paid employees.[15] Third, administrators disclose anonymized data to others within their organization.[16] Particularly within large organizations, data collectors may want to protect data subjects' privacy even from others in the organization.[17] For example, large banks may want to share some data with their marketing departments, but only after anonymizing it to protect customer privacy.

Lawrence Lessig's four regulators of behavior—norms and ethics, the market, architecture, and law—each compel administrators to anonymize.[18] Anonymization norms and ethics often operate through best practice documents that recommend anonymization as a technique for protecting privacy. For example, biomedical guidelines often recommend coding genetic

11. National Institutes of Health, HIPAA Privacy Rules for Researchers, http://privacyruleandresearch.nih.gov/faq.asp (last visited June 12, 2010).

12. *E.g.*, Posting of Susan Wojcicki, Vice President, Product Management to The Official Google Blog, Making Ads More Interesting, http://googleblog.blogspot.com/2009/03/making-ads-more-interesting.html (Mar. 11, 2009, 2:01 EST) (announcing a new Google initiative to tailor ads to "the types of sites you visit and the pages you view").

13. *E.g.*, *In re* Application of United States for an Order for Disclosure of Telecommunications Records and Authorizing the Use of a Pen Register and Trap and Trace, 405 F. Supp. 2d 435 (S.D.N.Y. 2005) (granting the government the authority to compel a provider to provide information suggesting the location of a customer's cell phone).

14. *See infra* Part I.B.1 (describing three public releases of databases).

15. *See* CLAY SHIRKY, HERE COMES EVERYBODY: THE POWER OF ORGANIZING WITHOUT ORGANIZATIONS (2008); JAMES SUROWIECKI, THE WISDOM OF CROWDS (2004).

16. *See* Posting of Philip Lenssen to Google Blogoscoped, Google-Internal Data Restrictions, http://blogoscoped.com/archive/2007-06-27-n27.html (June 27, 2007) (detailing how Google and Microsoft limit internal access to sensitive data).

17. *See id.*

18. *See* LAWRENCE LESSIG, CODE: VERSION 2.0, at 123 (2006) (listing four regulators of online behavior: markets, norms, laws, and architecture).

data—associating stored genes with nonidentifying numbers—to protect privacy.[19] Other guidelines recommend anonymization in contexts such as electronic commerce,[20] internet service provision,[21] data mining,[22] and national security data sharing.[23] Academic researchers rely heavily on anonymization to protect human research subjects, and their research guidelines recommend anonymization generally,[24] and specifically in education,[25] computer network monitoring,[26] and health studies.[27] Professional statisticians are duty-bound to anonymize data as a matter of professional ethics.[28]

Market pressures sometimes compel businesses to anonymize data. For example, companies like mint.com and wesabe.com provide web-based personal finance tracking and planning.[29] One way these companies add value is by aggregating and republishing data to help their customers compare their spending with that of similarly situated people.[30] To make customers comfortable with this type of data sharing, both mint.com and wesabe.com promise to anonymize data before sharing it.[31]

Architecture, defined in Lessig's sense as technological constraints,[32] often forces anonymization, or at least makes anonymization the default choice. As one example, whenever you visit a website, the distant computer with which you communicate—also known as the web server—records some information

---

19. Roberto Andorno, *Population Genetic Databases: A New Challenge to Human Rights*, *in* ETHICS AND LAW OF INTELLECTUAL PROPERTY 39 (Christian Lenk, Nils Hoppe & Roberto Andorno eds., 2007).

20. ALEX BERSON & LARRY DUBOV, MASTER DATA MANAGEMENT AND CUSTOMER DATA INTEGRATION FOR A GLOBAL ENTERPRISE 338–39 (2007).

21. *See infra* Part II.A.3.b.

22. G.K. GUPTA, INTRODUCTION TO DATA MINING WITH CASE STUDIES 432 (2006).

23. MARKLE FOUND. TASK FORCE, CREATING A TRUSTED NETWORK FOR HOMELAND SECURITY 144 (2003), *available at* http://www.markle.org/downloadable_assets/nstf_report2_full_report.pdf.

24. *See* THE SAGE ENCYCLOPEDIA OF QUALITATIVE RESEARCH METHODS 196 (Lisa M. Given ed., 2008) (entry for "Data Security").

25. LOUIS COHEN ET AL., RESEARCH METHODS IN EDUCATION 189 (2003).

26. *See* Ruoming Pang et al., *The Devil and Packet Trace Anonymization*, 36 COMP. COMM. REV. 29 (2006).

27. INST. OF MED., PROTECTING DATA PRIVACY IN HEALTH SERVICES RESEARCH 178 (2000).

28. European Union Article 29 Data Protection Working Party, *Opinion 4/2007 on the Concept of Personal Data*, 01248/07/EN WP 136, at 21 (June 20, 2007) [hereinafter 2007 Working Party Opinion], *available at* http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf.

29. *See* Eric Benderoff, *Spend and Save the Social Way—Personal Technology*, SEATTLE TIMES, Nov. 8, 2008, at A9.

30. *See* Carolyn Y. Johnson, *Online Social Networking Meets Personal Finance*, N.Y. TIMES, Aug. 7, 2007, *available at* http://www.nytimes.com/2007/08/07/technology/07iht-debt.1.7013213.html.

31. *See, e.g.*, Wesabe, Security and Privacy, http://www.wesabe.com/page/security (last visited June 12, 2010); Mint.com, How Mint Personal Finance Management Protects Your Financial Safety, http://www.mint.com/privacy (last visited June 12, 2010).

32. LESSIG, *supra* note 18, at 4.

about your visit into what is called a log file.[33] The vast majority of web servers collect much less than the maximum amount of information available about your visit, not due to the principled privacy convictions of their owners, but because the software saves only a limited amount of information by default.[34]

c. Faith in Anonymization

Many defend the privacy-protecting power of anonymization and hold it out as a best practice despite evidence to the contrary. In one best practices guide, the authors, after cursorily acknowledging concerns about the power of anonymization, conclude that, "[w]hile we recognize that [reidentification] is a remote possibility in some situations, in most cases genetic research data anonymization will help to ensure confidentiality."[35] Similarly, Google has said, "[i]t is difficult to guarantee complete anonymization, but we believe [Google's log file anonymization techniques] will make it very unlikely users could be identified."[36]

Government officials and policymakers embrace anonymization as well. Two influential data mining task forces have endorsed anonymization. In 2004, the Technology and Privacy Advisory Committee (TAPAC), a Defense Department–led group established in the wake of controversy over the government's Total Information Awareness program, produced an influential report about government data mining.[37] The report recommends anonymization "whenever practicable" and thus restricts all of its other recommendations only to databases that are not "known or reasonably likely to include personally identifiable information."[38]

Likewise, the Markle Foundation task force, which included among its members now–Attorney General Eric Holder, produced a similar report.[39] Like TAPAC, the Markle Foundation group concluded that "anonymizing technologies could be employed to allow analysts to perform link analysis among data sets without disclosing personally identifiable information . . . [so]

33. STEPHEN SPAINHOUR & ROBERT ECKSTEIN, WEBMASTER IN A NUTSHELL 458–59 (2002).

34. Apache, Apache HTTP Server Version 1.3 Log Files, http://httpd.apache.org/docs/1.3/logs.html (last visited June 12, 2010) (describing the default "common log format" which logs less information than the alternative "combined log format").

35. ADIL E. SHAMOO & DAVID B. RESNICK, RESPONSIBLE CONDUCT OF RESEARCH 302 (2009).

36. Chris Soghoian, *Debunking Google's Log Anonymization Propaganda, Surveillance State*, CNET NEWS, Sept. 11, 2008, http://news.cnet.com/8301-13739_3-10038963-46.html.

37. TECHNOLOGY & PRIVACY ADVISORY COMM., REPORT: SAFEGUARDING PRIVACY IN THE FIGHT AGAINST TERRORISM 35–36 (2004), *available at* http://www.cdt.org/security/usapatriot/20040300tapac.pdf.

38. *Id.* at 50 (Recommendation 2.2).

39. *See* MARKLE FOUND. TASK FORCE, *supra* note 23, at 34.

analysts can perform their jobs and search for suspicious patterns without the need to gain access to personal data until they make the requisite showing for disclosure."[40]

Many legal scholars share this faith in anonymization.[41] Ira Rubinstein, Ronald Lee, and Paul Schwartz state a "consensus view" that "[w]ith the goal of minimizing the amount of personal information revealed in the course of running pattern-based searches, the anonymization of data (such as names, addresses, and social security numbers) is essential."[42] Barbara Evans, a prominent medical privacy scholar, speaks about "anonymized" data "that have had patient identifiers completely and irrevocably removed before disclosure, such that future reidentification would be impossible."[43] Many other legal scholars have made similar claims premised on deep faith in robust anonymization.[44] The point is not to criticize or blame these people for trusting anonymization; as we will see, even computer scientists have been surprised by the success of recent attacks on anonymization.

2. Anonymization Techniques: The Release-and-Forget Model

How do people anonymize data? From among the scores of different anonymization techniques, I will focus on an important and large subset that I call release-and-forget anonymization.[45] As the name suggests, when a data administrator practices these techniques, she releases records—either publicly,

---

40. *Id.* at 34.

41. Regulators do too. *See infra* Part II.A (listing laws and regulations that assume robust anonymization).

42. Ira S. Rubinstein et al., *Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches*, 75 U. CHI. L. REV. 261, 266, 268 (2008).

43. Barbara J. Evans, *Congress' New Infrastructural Model of Medical Privacy*, 84 NOTRE DAME L. REV. 585, 619–20 (2009). Professor Evans has clarified that the quote did not reflect her personal opinions about the feasibility of definitive anonymization but rather reflected how the term 'anonymization' has commonly been understood by regulators and others in bioethics. Email From Barbara Evans, Assoc. Professor, Univ. of Houston Law Ctr., to Paul Ohm, Assoc. Professor, Univ. of Colorado Law Sch. (July 21, 2010) (on file with author).

44. *See, e.g.*, Fred H. Cate, *Government Data Mining: The Need for a Legal Framework*, 43 HARV. C.R.-C.L. L. REV. 435, 487 (2008); Matthew P. Gordon, *A Legal Duty to Disclose Individual Research Findings to Research Subjects?*, 64 FOOD & DRUG L.J. 225, 258–59 (2009); Bartha Maria Knoppers et al., *Ethical Issues in Secondary Uses of Human Biological Material From Mass Disasters*, 34 J.L. MED. & ETHICS 352, 353 (2006); Susan M. Wolf et al., *Managing Incidental Findings in Human Subjects Research: Analysis and Recommendations*, 36 J.L. MED. & ETHICS 219, 226–27 (2008); Irfan Tukdi, Comment, *Transatlantic Turbulence: The Passenger Name Record Conflict*, 45 HOUS. L. REV. 587, 618–19 (2008).

45. Other means of making data more anonymous include releasing only aggregated statistics; interactive techniques, in which administrators answer directed questions on behalf of researchers, instead of releasing data in its entirety; and "differential privacy" techniques, which protect privacy by adding carefully calibrated noise to the data. *See* discussion *infra* Part III.B.2.

privately to a third party, or internally within her own organization—and then she forgets, meaning she makes no attempt to track what happens to the records after release. Rather than blithely put her data subjects at risk, before she releases, she modifies some of the information.

I focus on release-and-forget anonymization for two reasons. First, these techniques are widespread.[46] Because they promise privacy while allowing the broad dissemination of data, they give data administrators everything they want without any compromises, and data administrators have embraced them.[47] Second, these techniques are often flawed. Many of the recent advances in the science of reidentification target release-and-forget anonymization in particular.[48]

Consider some common release-and-forget techniques.[49] First, we need a sample database to anonymize, a simplified and hypothetical model of a hospital's database for tracking visits and complaints:[50]

TABLE 1: Original (Nonanonymized) Data

Name	Race	Birth Date	Sex	ZIP Code	Complaint
Sean	Black	9/20/1965	Male	02141	Short of breath
Daniel	Black	2/14/1965	Male	02141	Chest pain
Kate	Black	10/23/1965	Female	02138	Painful eye
Marion	Black	8/24/1965	Female	02138	Wheezing
Helen	Black	11/7/1964	Female	02138	Aching joints
Reese	Black	12/1/1964	Female	02138	Chest pain
Forest	White	10/23/1964	Male	02138	Short of breath
Hilary	White	3/15/1965	Female	02139	Hypertension
Philip	White	8/13/1964	Male	02139	Aching joints
Jamie	White	5/5/1964	Male	02139	Fever
Sean	White	2/13/1967	Male	02138	Vomiting
Adrien	White	3/21/1967	Male	02138	Back pain

46. *See* Laks V.S. Lakshmanan & Raymond T. Ng, *On Disclosure Risk Analysis of Anonymized Itemsets in the Presence of Prior Knowledge*, 2 ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA 13, 13:2 (2008) ("Among the well-known transformation techniques, *anonymization* is arguably the most common.").

47. *Id.* ("Compared with other transformation techniques, anonymization is simple to carry out, as mapping objects back and forth is easy.").

48. *See* Justin Brickell & Vitaly Shmatikov, *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*, *in* 2008 KNOWLEDGE DISCOVERY & DATA MINING CONF. 70, 70.

49. The following discussion is only a survey; it will make an expert of no one.

50. All of the hypothetical data in this table aside from the "Name" column comes from a paper by Latanya Sweeney. Sweeney, *supra* note 8, at 567 fig.4. Where the first names come from is left as an exercise for the reader.

Using standard terminology, we call this collection of data a table; each row is a row or record; each column is a column, field, or attribute, identified by a label (in bold) called a field name or attribute name; each record has a particular value for a given attribute.[51]

To protect the privacy of the people in this table, the hospital database administrator will take the following steps before releasing this data:

Singling Out Identifying Information: First, the administrator will single out any fields she thinks one can use to identify individuals. Often, she will single out not only well-known identifiers like name and social security number, but combinations of fields that when considered together might link a record in the table to a patient's identity.[52] Sometimes an administrator will select the potentially identifying fields herself, either intuitively (by isolating types of data that seem identifying) or analytically (by looking for uniqueness in the particular data). For example, no two people in our database share a birth date, so the administrator must treat birth date as an identifier.[53] If she did not, then anyone who knew Forest's birth date (and who knew Forest had been admitted to the hospital) would be able to find Forest in the anonymized data.[54]

In other cases, an administrator will look to another source—such as a statistical study, company policy, or government regulation—to decide whether or not to treat a particular field as identifying. In this case, assume the administrator decides, based on one of these sources, to treat the following four fields as potential identifiers: name, birth date, sex, and ZIP code.[55]

Suppression: Next, the administrator will modify the identifying fields. She might suppress them, removing the fields from the table altogether.[56] In our example, the administrator might delete all four potential identifiers, producing this table:

51. GAVIN POWELL, BEGINNING DATABASE DESIGN 38–41 (2005).

52. Claudio Bettini et al., *The Role of Quasi-Identifiers in k-Anonymity Revisited* (DICo Univ. Milan Tech. Rep. RT-11-06, July 2006).

53. *See id.* Because these sorts of identifiers do not link directly to identity, researchers sometimes refer to them as quasi-identifiers.

54. That large numbers of people could know Forest's birth date is far from an idle worry. Today, more than ever, people are sharing this kind of information widely. For example, "at least 10 million U.S. residents make publicly available or inferable their birthday information on their [social networking] online profiles." Alessandro Acquisti & Ralph Gross, *SSN Study-FAQ*, http://www.heinz.cmu.edu/~acquisti/ssnstudy (last visited June 12, 2010).

55. *See infra* Part I.B.1.b (discussing research about using the combination of ZIP code, birth date, and sex as an identifier).

56. Sweeney, *supra* note 8, at 3.

TABLE 2: Suppressing Four Identifier Fields

Race	Complaint
Black	Short of breath
Black	Chest pain
Black	Painful eye
Black	Wheezing
Black	Aching joints
Black	Chest pain
White	Short of breath
White	Hypertension
White	Aching joints
White	Fever
White	Vomiting
White	Back pain

Here we first encounter a fundamental tension. On the one hand, with this version of the data, we should worry little about privacy; even if one knows Forest's birth date, sex, ZIP code, and race, one still cannot learn Forest's complaint. On the other hand, aggressive suppression has rendered this data almost useless for research.[57] Although a researcher can use the remaining data to track the incidence of diseases by race, because age, sex, and residence have been removed, the researcher will not be able to draw many other interesting and useful conclusions.

Generalization: To better strike the balance between utility and privacy, the anonymizer might generalize rather than suppress identifiers.[58] This means she will alter rather than delete identifier values to increase privacy while preserving utility. For example, the anonymizer may choose to suppress the name field, generalize the birth date to only the year of birth, and generalize ZIP codes by retaining only the first three digits.[59] The resulting data would look like this:

57. *See infra* Part III.B.1 (discussing the relationship between utility and privacy).

58. Sweeney, *supra* note 8, at 3.

59. Under the HIPAA Privacy Rule, these three changes would qualify the resulting table as deidentified health information. *See* U.S. Health & Human Services, Standards for Privacy of Individually Identifiable Health Information, 45 C.F.R. §§ 160, 164 (2009). For more on HIPAA and the Privacy Rule, see *infra* Part II.A.3.a.

TABLE 3: Generalized

Race	Birth Year	Sex	ZIP Code*	Complaint
Black	**1965**	Male	**021***	Short of breath
Black	**1965**	Male	**021***	Chest pain
Black	**1965**	Female	**021***	Painful eye
Black	**1965**	Female	**021***	Wheezing
Black	**1964**	Female	**021***	Aching joints
Black	**1964**	Female	**021***	Chest pain
White	**1964**	Male	**021***	Short of breath
White	**1965**	Female	**021***	Hypertension
White	**1964**	Male	**021***	Aching joints
White	**1964**	Male	**021***	Fever
White	**1967**	Male	**021***	Vomiting
White	**1967**	Male	**021***	Back pain

Now, even someone who knows Forest's birth date, ZIP code, sex, and race will have trouble plucking out Forest's specific complaint. The records in this generalized data (Table 3) are more difficult to reidentify than they were in the original data (Table 1), but researchers will find this data much more useful than the suppressed data (Table 2).

Aggregation: Finally, to better understand what qualifies as release-and-forget anonymization, consider a commonly used technique that does not obey release-and-forget. Quite often, an analyst needs only summary statistics, not raw data. For decades, statisticians have investigated how to release aggregate statistics while protecting data subjects from reidentification.[60] Thus, if researchers only need to know how many men complained of shortness of breath, data administrators could release this:

TABLE 4: Aggregate Statistic

Men Short of Breath	2

60. *E.g.*, Nabil R. Adam & John C. Wortmann, *Security-Control Methods for Statistical Databases: A Comparative Study*, 21 ACM COMPUTING SURVEYS 515 (1989); Tore Dalenius, *Towards a Methodology for Statistical Disclosure Control*, 15 STATISTISK TIDSKRIFT 429 (1977) (Swed.); I.P. Fellegi, *On the Question of Statistical Confidentiality*, 67 J. AM. STAT. ASS'N 7 (1972).

As it happens, Forest is one of the two men described by this statistic—he complained about shortness of breath—but without a lot of additional information, one would never know. His privacy is secure.[61]

Privacy lawyers tend to refer to release-and-forget anonymization techniques using two other names: deidentification[62] and the removal of personally identifiable information (PII).[63] Deidentification has taken on special importance in the health privacy context. Regulations implementing the privacy provisions of the Health Insurance Portability and Accountability Act (HIPAA) expressly use the term, exempting health providers and researchers who deidentify data before releasing it from all of HIPAA's many onerous privacy requirements.[64]

B. The Present and Future: Easy Reidentification

Until a decade ago, the robust anonymization assumption worked well for everybody involved. Data administrators could protect privacy when sharing data with third parties; data subjects could rest assured that their secrets would remain private; legislators could balance privacy and other interests (such as the advancement of knowledge) by deregulating the trade in anonymized records;[65] and regulators could easily divide data handlers into two groups: the responsible (those who anonymized) and the irresponsible (those who did not).

About fifteen years ago, researchers started to chip away at the robust anonymization assumption, the foundation upon which this state of affairs has been built. Recently, however, they have done more than chip away; they have essentially blown it up, casting serious doubt on the power of anonymization, proving its theoretical limits and establishing what I call the easy reidentification result. This is not to say that all anonymization techniques fail to protect privacy—some techniques are very difficult to reverse—but researchers have learned more than enough already for us to reject anonymization as a privacy-providing panacea.

---

61. For additional discussion of privacy techniques other than release-and-forget, see *infra* Part III.B.2.

62. National Institutes of Health, *De-identifying Protected Health Information Under the Privacy Rule*, http://privacyruleandresearch.nih.gov/pr_08.asp (last visited June 12, 2010).

63. ERIKA MCCALLISTER ET AL., NAT'L INST. OF STANDARDS & TECH., SPECIAL PUB. NO. 800-122, GUIDE TO PROTECTING THE CONFIDENTIALITY OF PERSONALLY IDENTIFIABLE INFORMATION (PII) (2010), *available at* http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf.

64. 45 C.F.R. §§ 164.502(d)(2), 164.514(a)–(b) (2009). *See infra* Part II.A.3.a.

65. *See infra* II.A.

1. How Three Anonymized Databases Were Undone

Consider three recent, spectacular failures of anonymization. In each case, a sophisticated entity placed unjustified faith in weak, release-and-forget anonymization. These stories, which I will use as examples throughout this Article, provide two important lessons: They demonstrate the pervasiveness of release-and-forget anonymization even among supposedly sophisticated data administrators, and they demonstrate the peril of this kind of anonymization in light of recent advances in reidentification.

a. The AOL Data Release

On August 3, 2006, America Online (AOL) announced a new initiative called "AOL Research."[66] To "embrac[e] the vision of an open research community," AOL Research publicly posted to a website twenty million search queries for 650,000 users of AOL's search engine, summarizing three months of activity.[67] Researchers of internet behavior rejoiced to receive this treasure trove of information, the kind of information that is usually treated by search engines as a closely guarded secret.[68] The euphoria was short-lived, however, as AOL and the rest of the world soon learned that search engine queries are windows to the soul.

Before releasing the data to the public, AOL had tried to anonymize it to protect privacy. It suppressed any obviously identifying information such as AOL username and IP address in the released data.[69] In order to preserve the usefulness of the data for research, however, it replaced these identifiers with unique identification numbers that allowed researchers to correlate different searches to individual users.[70]

In the days following the release, bloggers pored through the data spotlighting repeatedly the nature and extent of the privacy breach. These bloggers chased two different prizes, either attempting to identify users or

66. Posting of Abdur Chowdhury, cabdur@aol.com, to SIGIR-IRList, irlist-editor@acm.org, http://sifaka.cs.uiuc.edu/xshen/aol/20060803_SIG-IRListEmail.txt (last visited July 19, 2010).

67. *Id.* Others have reported that the data contained thirty-six million entries. Paul Boutin, *You Are What You Search*, SLATE, Aug. 11, 2006, http://www.slate.com/id/2147590.

68. *See* Katie Hafner, *Researchers Yearn to Use AOL Logs, but They Hesitate*, N.Y. TIMES, Aug 23, 2006, at C1 (describing the difficulty that academic researchers experience accessing raw search data).

69. *See* Michael Barbaro & Tom Zeller, Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES, Aug. 9, 2006, at A1. IP addresses, discussed *infra* in Part II.A.3.b, are numbers that identify computers on the internet and can be used to track internet activity.

70. Barbaro & Zeller, Jr., *supra* note 69.

"hunt[ing] for particularly entertaining or shocking search histories."[71] Thanks to this blogging and subsequent news reporting, certain user identification numbers have become sad little badges of infamy, associated with pitiful or chilling stories. User "No. 3505202 ask[ed] about 'depression and medical leave.' No. 7268042 type[d] 'fear that spouse contemplating cheating.'"[72] User 17556639 searched for "how to kill your wife" followed by a string of searches for things like "pictures of dead people" and "car crash photo."[73]

While most of the blogosphere quickly and roundly condemned AOL,[74] a few bloggers argued that the released data, while titillating, did not violate privacy because nobody had linked actual individuals with their anonymized queries.[75] This argument was quickly silenced by *New York Times* reporters Michael Barbaro and Tom Zeller, who recognized clues to User 4417749's identity in queries such as "'landscapers in Lilburn, Ga,' several people with the last name Arnold and 'homes sold in shadow lake subdivision gwinnett county georgia.'"[76] They quickly tracked down Thelma Arnold, a sixty-two-year-old widow from Lilburn, Georgia who acknowledged that she had authored the searches, including some mildly embarrassing queries such as "numb fingers," "60 single men," and "dog that urinates on everything."[77]

The fallout was swift and crushing. AOL fired the researcher who released the data and also his supervisor.[78] Chief Technology Officer Maureen Govern resigned.[79] The fledgling AOL Research division has been silenced, and a year after the incident, the group still had no working website.[80]

---

71. *Id.* These twin goals demonstrate an important information dichotomy revisited later: When someone talks about the sensitivity of data, they may mean that the information can cause harm if disclosed, or they may mean that the information can be used to link anonymized information to identity. As we will see, regulators often misunderstand the difference between these two classes of information. *See infra* Part II.A.

72. *See* Barbaro & Zeller, Jr., *supra* note 69.

73. Markus Frind, *AOL Search Data Shows Users Planning to Commit Murder, Paradigm Shift Blog* (Aug. 7, 2006), http://plentyoffish.wordpress.com/2006/08/07/aol-search-data-shows-users-planning-to-commit-murder.

74. *See, e.g.*, Posting of Michael Arrington to TechCrunch, *AOL Proudly Releases Massive Amounts of Private Data* (Aug. 6, 2006), http://www.techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data ("The utter stupidity of this is staggering.").

75. Greg Linden, for example, complained that "no one actually has come up with an example where someone could be identified. Just the theoretical possibility is enough to create a privacy firestorm in some people's minds." Greg Linden, *A Chance to Play With Big Data: Geeking With Greg*, http://glinden.blogspot.com/2006/08/chance-to-play-with-big-data.html (Aug. 4, 2006, 19:53 PST).

76. Barbaro & Zeller, Jr., *supra* note 69.

77. *Id.*

78. Tom Zeller, Jr., *AOL Executive Quits After Posting of Search Data*, N.Y. TIMES, Aug. 22, 2006, http://www.nytimes.com/2006/08/22/technology/22iht-aol.2558731.html.

79. *Id.*

80. Chris Soghoian, *AOL, Netflix and the End of Open Access to Research Data, Surveillance State*, CNET NEWS, Nov. 30, 2007, http://news.cnet.com/8301-13739_3-9826608-46.html.

b. ZIP, Sex, and Birth Date

Recall from the Introduction the study by Latanya Sweeney, professor of computer science, who crunched 1990 census data and discovered that 87.1 percent of people in the United States were uniquely identified by their combined five-digit ZIP code, birth date (including year), and sex.[81] According to her study, even less-specific information can often reveal identity, as 53 percent of American citizens are uniquely identified by their city, birth date, and sex, and 18 percent by their county, birth date, and sex.[82]

Like the reporters who discovered Thelma Arnold, Dr. Sweeney offered a hyper-salient example to drive home the power (and the threat) of reidentification techniques. In Massachusetts, a government agency called the Group Insurance Commission (GIC) purchased health insurance for state employees.[83] At some point in the mid-1990s, GIC decided to release records summarizing every state employee's hospital visits at no cost to any researcher who requested them.[84] By removing fields containing name, address, social security number, and other "explicit identifiers," GIC assumed it had protected patient privacy, despite the fact that "nearly one hundred attributes per" patient and hospital visit were still included, including the critical trio of ZIP code, birth date, and sex.[85]

At the time that GIC released the data, William Weld, then–Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers.[86] In response, then–graduate student Sweeney started hunting for the Governor's hospital records in the GIC data.[87] She knew that Governor Weld resided in Cambridge, Massachusetts, a city of fifty-four thousand residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge—a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor

81. Sweeney, *supra* note 4. A subsequent study placed the number at 61 percent (for 1990 census data) and 63 percent (for 2000 census data). Golle, *supra* note 4, at 1.

82. Sweeney, *supra* note 4.

83. Massachusetts Executive Office for Administration and Finance, *Who is the GIC?*, http://mass.gov/gic (follow "Who is the GIC?" hyperlink) (last visited June 15, 2010).

84. *Recommendations to Identify and Combat Privacy Problems in the Commonwealth: Hearing on H.R. 351 Before the H. Select Comm. on Information Security*, 189th Sess. (Pa. 2005) (statement of Latanya Sweeney, Associate Professor, Carnegie Mellon University), *available at* http://dataprivacylab.org/dataprivacy/talks/Flick-05-10.html.

85. *Id.*

86. Henry T. Greely, *The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks*, 8 ANN. REV. GENOMICS & HUM. GENETICS 343, 352 (2007).

87. *Id.*

Weld with ease. Only six people in Cambridge shared his birth date; only three were men, and of the three, only he lived in his ZIP code.[88] In a theatrical flourish, Dr. Sweeney sent the governor's health records (including diagnoses and prescriptions) to his office.[89]

c. The Netflix Prize Data Study

On October 2, 2006, about two months after the AOL debacle, Netflix, the "world's largest online movie rental service," publicly released one hundred million records revealing how nearly a half-million of its users had rated movies from December 1999 to December 2005.[90] In each record, Netflix disclosed the movie rated, the rating assigned (from one to five stars), and the date of the rating.[91] Like AOL and GIC, Netflix first anonymized the records, removing identifying information like usernames, but assigning a unique user identifier to preserve rating-to-rating continuity.[92] Thus, researchers could tell that user 1337 had rated *Gattaca* a 4 on March 3, 2003, and *Minority Report* a 5 on November 10, 2003.

Unlike AOL, Netflix had a specific profit motive for releasing these records.[93] Netflix thrives by being able to make accurate movie recommendations; if Netflix knows, for example, that people who liked *Gattaca* will also like *The Lives of Others*, it can make recommendations that keep its customers coming back to the website.

To improve its recommendations, Netflix released the hundred million records to launch what it called the "Netflix Prize," a prize that took almost three years to claim.[94] The first team that used the data to significantly improve on Netflix's recommendation algorithm would win one million dollars.[95] As with the AOL release, researchers have hailed the Netflix Prize data release as a great boon for research, and many have used the competition to refine or develop important statistical theories.[96]

88. Sweeney, *supra* note 4.
89. Greely, *supra* note 86.
90. The Netflix Prize Rules, http://www.netflixprize.com/rules (last visited June 12, 2010).
91. *Id*.
92. Netflix Prize: FAQ, http://www.netflixprize.com/faq (last visited June 12, 2010) (answering the question, "Is there any customer information in the dataset that should be kept private?").
93. *See* Clive Thompson, *If You Liked This, You're Sure to Love That*, N.Y. TIMES MAG., Nov. 23, 2008, at 74, *available at* http://www.nytimes.com/2008/11/23/magazine/23Netflix-t.html.
94. Posting of Steve Lohr, *Netflix Challenge Ends, but Winner is in Doubt*, N.Y. TIMES BITS BLOG, http://bits.blogs.nytimes.com/2009/07/27/netflix-challenge-ends-but-winner-is-in-doubt (July 27, 2009, 16:59 EST).
95. *See* The Netflix Prize Rules, *supra* note 90.
96. *See* Thompson, *supra* note 93.

Two weeks after the data release, researchers from the University of Texas, Arvind Narayanan and Professor Vitaly Shmatikov, announced that "an attacker who knows only a little bit about an individual subscriber can easily identify this subscriber's record if it is present in the [Netflix Prize] dataset, or, at the very least, identify a small set of records which include the subscriber's record."[97] In other words, it is surprisingly easy to reidentify people in the database and thus discover all of the movies they have rated with only a little outside knowledge about their movie-watching preferences.

The resulting research paper is brimming with startling examples of the ease with which someone could reidentify people in the database, and has been celebrated and cited as surprising and novel to computer scientists.[98] If an adversary—the term used by computer scientists[99]—knows the precise ratings a person in the database has assigned to six obscure movies,[100] and nothing else, he will be able to identify that person 84 percent of the time.[101] If he knows approximately when (give or take two weeks) a person in the database has rated six movies, whether or not they are obscure, he can identify the person 99 percent of the time.[102] In fact, knowing when ratings were assigned turns out to be so powerful that knowing only two movies a rating user has viewed (with the precise ratings and the rating dates give or take three days), an adversary can reidentify 68 percent of the users.[103]

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.

To turn these abstract results into concrete examples, Narayanan and Shmatikov compared the Netflix rating data to similar data from the Internet

---

97. Arvind Narayanan & Vitaly Shmatikov, *How to Break the Anonymity of the Netflix Prize Dataset*, ARVIX, Oct. 16, 2006, at 1, http://arxiv.org/abs/cs/0610105v1 (v.1) [hereinafter *Netflix Prize v1*]. Narayanan and Shmatikov eventually published the results in 2008. *Netflix Prize Study*, *supra* note 5.

98. In 2008, the paper was awarded the "Award for Outstanding Research in Privacy Enhancing Technologies" or PET Award, given jointly by Microsoft and the Privacy Commissioner of Ontario, Canada. Press Release, EMEA Press Ctr., Microsoft, Privacy to the Test—Exploring the Limits of Online Anonymity and Accountability (July 23, 2008), http://www.microsoft.com/emea/presscentre/pressreleases/23072008_PETSFS.mspx. *E.g.*, Cynthia Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *in* PRIVACY, SECURITY, AND TRUST IN KDD 1, 2 (2008), *available at* http://www.springerlink.com/content/85g8155l38612w06/fulltext.pdf.

99. *See infra* Part I.B.2.a.

100. By obscure movie, I mean a movie outside the top five hundred movies rated in the database, ranked by number of ratings given. *See generally Netflix Prize Study*, *supra* note 5.

101. *Id.* at 121, 122 fig.8. The authors emphasize that this result would apply to most of the rating users, as 90 percent of them rated five or more obscure movies and 80 percent rated ten or more obscure movies. *Id.* at 121 tbl.

102. *Id.* at 121, 120 fig.4.

103. *Id.*

Movie Database (IMDb),[104] a movie-related website that also gives users the chance to rate movies. Unlike Netflix, IMDb posts these ratings publicly on its website, as Amazon does with user-submitted book ratings.

Narayanan and Shmatikov obtained ratings for fifty IMDb users.[105] From this tiny sample,[106] they found two users who were identifiable, to a statistical near-certainty, in the Netflix database.[107] Because neither database comprised a perfect subset of the other, one could learn things from Netflix unknowable only from IMDb, and vice versa,[108] including some things these users probably did not want revealed. For example, the authors listed movies viewed by one user that suggested facts about his or her politics ("Fahrenheit 9/11"), religious views ("Jesus of Nazareth"), and attitudes toward gay people ("Queer as Folk").[109]

Soon after it awarded the first Netflix Prize, the company announced that it would launch a second contest, one involving "demographic and behavioral data . . . includ[ing] information about renters' ages, gender, ZIP codes, genre ratings, and previously chosen movies."[110] In late 2009, a few Netflix customers brought a class action lawsuit against the company for privacy violations stemming from the release of their information through the Netflix Prize.[111] The suit alleged violations of various state and federal privacy laws.[112] A few months later, after the FTC became involved, Netflix announced that it had settled the suit and shelved plans for the second contest.[113]

104. Internet Movie Database, http://www.imdb.com (last visited June 12, 2010).

105. Ideally, the authors would have imported the entire IMDb ratings database to see how many people they could identify in the Netflix data. The authors were afraid, however, that the IMDb terms of service prohibited this. *Netflix Prize Study*, *supra* note 5, at 122. As of Feb. 11, 2009, the IMDb terms of service prohibited, among other things, "data mining, robots, screen scraping, or similar data gathering and extraction tools." Internet Movie Database, IMDb Copyright and Conditions of Use, http://www.imdb.com/help/show_article?conditions (last visited June 12, 2010).

106. IMDb reports that 57 million users visit its site each month. Internet Movie Database, IMDb History, http://www.imdb.com/help/show_leaf?history (last visited June 12, 2010).

107. *Netflix Prize Study*, *supra* note 5, at 123.

108. *Id.*

109. *Id.*

110. Posting of Steve Lohr, *Netflix Awards $1 Million Prize and Starts a New Contest*, N.Y. TIMES BITS BLOG, http://bits.blogs.nytimes.com/2009/09/21/netflix-awards-1-million-prize-and-starts-a-new-contest (Sep. 21, 2009, 10:15 EST).

111. Posting of Ryan Singel, *Netflix Spilled Your* Brokeback Mountain *Secret, Lawsuit Claims*, WIRED THREAT LEVEL BLOG, http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit (Dec. 17, 2009, 16:29 EST).

112. *Id.*

113. Posting of Steve Lohr, *Netflix Cancels Contest Plans and Settles Suit*, N.Y. TIMES BITS BLOG, http://bits.blogs.nytimes.com/2010/03/12/netflix-cancels-contest-plans-and-settles-suit (Mar. 12, 2010, 2:46 PM EST).

2. Reidentification Techniques

How did Sweeney discover William Weld's diagnoses? How did Barbaro and Zeller find Thelma Arnold? How did Narayanan and Shmatikov reidentify the people in the Netflix Prize dataset? Each researcher combined two sets of data—each of which provided partial answers to the question "who does this data describe?"—and discovered that the combined data answered (or nearly answered) the question.

Even though administrators had removed any data fields they thought might uniquely identify individuals, researchers in each of the three cases unlocked identity by discovering pockets of surprising uniqueness remaining in the data. Just as human fingerprints left at a crime scene can uniquely identify a single person and link that person with "anonymous" information, so too do data subjects generate "data fingerprints"—combinations of values of data shared by nobody else in their table.[114]

Of course, researchers have long understood the basic intuition behind a data fingerprint; this intuition lay at the heart of endless debates about personally identifiable information (PII). What has startled observers about the new results, however, is that researchers have found data fingerprints in non-PII data, with much greater ease than most would have predicted. It is this element of surprise that has so disrupted the status quo. Sweeney realized the surprising uniqueness of ZIP codes, birth dates, and sex in the U.S. population; Barbaro and Zeller relied upon the uniqueness of a person's search queries; and Narayanan and Shmatikov unearthed the surprising uniqueness of the set of movies a person had seen and rated. These results suggest that maybe everything is PII to one who has access to the right outside information. Although many of the details and formal proofs of this work are beyond the scope of this Article, consider a few aspects of the science that are relevant to law and policy.

a. The Adversary

Computer scientists model anonymization and reidentification as an adversarial game, with anonymization simply an opening move.[115] They call the

114. *See* BBN Tech., *Anonymization & Deidentification*, http://www.bbn.com/technology/hci/security/anon (last visited June 12, 2010) (referring to services to remove "'fingerprints' in the data").

115. *See* Irit Dinur & Kobbi Nissim, *Revealing Information While Preserving Privacy*, *in* PROC. 22ND ACM SYMP. ON PRINCIPLES DATABASE SYS. 202, 203 (2003), *available at* http://portal.acm.org/citation.cfm?id=773173.

person trying to reidentify the data the "adversary."[116] They seem not to moralize the adversary, making no assumptions about whether he or she wants to reidentify for good or ill. The defining feature of the adversary seems to be that he or she is, no surprise, adversarial—motivated to do something the data administrator wishes not to happen.

Who are these potential adversaries who might have a motive to reidentify? Narayanan and Shmatikov suggest "stalkers, investigators, nosy colleagues, employers, or neighbors."[117] To this list we can add the police, national security analysts, advertisers, and anyone else interested in associating individuals with data.

b. Outside Information

Once an adversary finds a unique data fingerprint, he can link that data to outside information, sometimes called auxiliary information.[118] Many anonymization techniques would be perfect, if only the adversary knew nothing else about people in the world. In reality, of course, the world is awash in data about people, with new databases created every day. Adversaries combine anonymized data with outside information to pry out obscured identities.

Computer scientists make one appropriately conservative assumption about outside information that regulators should adopt: We cannot predict the type and amount of outside information the adversary can access.[119] It is naïve to assume that the adversary will be unable to find the particular piece of data needed to unlock anonymized data.[120] In computer security, this discredited attitude is called "security through obscurity."[121] Not only do reidentification scientists spurn security through obscurity, but they often assume that the adversary possesses the exact piece of data—if it exists—needed to unlock anonymized identities, in order to design responses that protect identity even in this worst case.[122]

116. *Id.*

117. Arvind Narayanan & Vitaly Shmatikov, *De-Anonymizing Social Networks*, *in* PROC. 2009 30TH IEEE SYMP. ON SECURITY & PRIVACY 173, 203 [hereinafter *De-Anonymizing Social Networks*] (for a draft version of this article that includes unpublished appendices, see Narayanan & Shmatikov, *infra* note 169).

118. *See Netflix Prize Study*, *supra* note 5, at 112.

119. *Id.*

120. *Id.*

121. SIMSON GARFINKEL ET AL., PRACTICAL UNIX AND INTERNET SECURITY 61 (2003) (describing "[t]he problem with security through obscurity").

122. *Cf.* Cynthia Dwork, *Differential Privacy*, *in* AUTOMATA, LANGUAGES AND PROGRAMMING, 33RD INT'L COLLOQUIUM PROC. PART II 1, 2 (2006), *available at* http://www.springerlink.com/content/383p21xk13841688/fulltext.pdf.

It seems wise to adopt this aggressively pessimistic assumption of perfect outside information given the avalanche of information now available on the internet[123] and, in particular, the rise of blogs and social networks. Never before in human history has it been so easy to peer into the private diaries of so many people.[124] Alessandro Acquisti and Ralph Gross—researchers who developed an efficient algorithm for using public data to guess people's social security numbers[125]—call this the "age of self-revelation."[126]

As only one example among many, in early 2009, many Facebook users began posting lists called "25 random things about me."[127] The implicit point of the exercise was to bare one's soul—at least a little—by revealing secrets about oneself that friends would not already know.[128] "25 random things about me" acts like a reidentification virus[129] because it elicits a vast amount of secret information in a concise, digital format. This is but one example of the rich outside information available on social networking websites. It is no surprise that several researchers have already reidentified people in anonymized social networking data.[130]

c. The Basic Principle: Of Crossed Hands and Inner Joins

One computer security expert summarized the entire field of reidentification to me with a simple motion: He folded his hands together, interleaving his fingers, like a parishioner about to pray. This simple mental image nicely summarizes the basic reidentification operation. If you imagine that your left hand is anonymized data, your right hand is outside information, and your interleaved fingers are places where information from the left matches the right, this image basically captures how reidentification is achieved.

---

123. *See* Lakshmanan & Ng, *supra* note 46, at 13:3 ("The assumption that there is no partial [outside] information out there is simply unrealistic in this Internet era.").

124. *Cf. De-Anonymizing Social Networks*, *supra* note 117, at 173–74 (describing sharing of information obtained from social networks).

125. Alesandro Acquisti & Ralph Gross, *Predicting Social Security Numbers from Public Data*, 106 NAT'L ACAD. SCI. 10975 (2009).

126. Acquisti & Gross, *supra* note 54.

127. Douglas Quenqua, *Ah, Yes, More About Me? Here are '25 Random Things'*, N.Y. TIMES, Feb. 4, 2009, at E6.

128. *See id*.

129. *E.g.*, Michael Kruse, *25 Random Things About Me to Keep You Caring*, ST. PETERSBURG TIMES, Feb. 23, 2009, *available at* http://www.tampabay.com/features/humaninterest/article978293.ece.

130. *De-Anonymizing Social Networks*, *supra* note 117, at 177; *see also* Lars Backstrom, Cynthia Dwork & Jon Kleinberg, *Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, *in* 16TH INT'L WORLD WIDE WEB CONFERENCE PROC. 181 (2007), *available at* http://portal.acm.org/citation.cfm?id=1242598.

Database administrators call the hand-folding operation an "inner join."[131] An inner join is an operation combining two database tables, connecting rows from one to rows from the other by matching shared information.[132] When the rows in the tables represent people, an inner join assumes that rows in which critical fields match refer to the same person, and can be combined into one row in the output table.[133] For example, if an adversary has one table that looks like this:

TABLE 5: Anonymized Database

Race	Birth Date	Sex	ZIP Code	Complaint
Black	9/20/1965	Male	02141	Short of breath
Black	2/14/1965	Male	02141	Chest pain
Black	10/23/1965	Female	02138	Painful eye
Black	8/24/1965	Female	02138	Wheezing
Black	11/7/1964	Female	02138	Aching joints
Black	12/1/1964	Female	02138	Chest pain
White	10/23/1964	Male	02138	Short of breath
White	3/15/1965	Female	02139	Hypertension
White	8/13/1964	Male	02139	Aching joints
White	5/5/1964	Male	02139	Fever
White	2/13/1967	Male	02138	Vomiting
White	3/21/1967	Male	02138	Back pain

131. Indeed, in common database systems "INNER JOIN" is the command used to perform such an operation. *See, e.g.*, ALAN BEAULIEU, LEARNING SQL 77 (2005); ANDY OPPEL & ROBERT SHELDON, SQL: A BEGINNER'S GUIDE 264 (2009); ALLEN G. TAYLOR, SQL ALL-IN-ONE DESK REFERENCE FOR DUMMIES 309 (2007); PAUL WILTON & JOHN COLBY, BEGINNING SQL 90–93 (2005).

132. *See* BEAULIEU, *supra note* 131.

133. *See id.* This simple example necessarily masks some complexity. For example, reidentifiers must contend with noisy data—errors that cause false positives and false negatives in the inner join. They use probability theory to spot both of these kinds of errors. *See Netflix Prize Study*, *supra* note 5, at 120.

and a separate table that looks like this:

TABLE 6: Database Including PII

Name	Birth Date	Sex	ZIP Code	Smoker?
Daniel	2/14/1965	Male	02141	Yes
Forest	10/23/1964	Male	02138	Yes
Helen	11/7/1964	Feale	02138	No
Hilary	3/15/1965	Female	02139	No
Kate	10/23/1965	Female	02138	No
Marion	8/24/1965	Female	02138	Yes

and she performs an inner join on the birth date, sex, and ZIP code columns, she would produce this:

TABLE 7: Inner Join of Tables 5 and 6 on Birth Date/ZIP/Sex

Name	Race	Birth Date	Sex	ZIP Code	Complaint	Smoker?
Daniel	Black	2/14/1965	Male	02141	Chest pain	Yes
Kate	Black	10/23/1965	Female	02138	Painful eye	No
Marion	Black	8/24/1965	Female	02138	Wheezing	Yes
Helen	Black	11/7/1964	Female	02138	Aching joints	No
Forest	White	10/23/1964	Male	02138	Short of breath	Yes
Hilary	White	3/15/1965	Female	02139	Hypertension	No

Notice that with the two joined tables, the sum of the information is greater than the parts. From the first table alone, the adversary did not know that the white male complaining of shortness of breath was Forest, nor did he know that the person was a smoker. From the second table alone, the adversary knew nothing about Forest's visit to the hospital. After the inner join, the adversary knows all of this.

3. Responding to Objections

In the rest of this Article, I draw many lessons from the three stories presented above and use these lessons to call for aggressive regulatory responses to the failure of anonymization. I anticipate, and in some cases I have confronted, several objections to these interpretations and prescriptions that deserve responses.

a. No Harm, No Foul

The three stories above demonstrate well the power of reidentification, but they do not demonstrate how reidentification can be used to harm people. The researchers described are professional journalists or academics, and ethical rules and good moral judgment limited the harm they caused. But do not be misled if the results of these studies seem benign. In Part III, I show how the techniques used in these studies can lead to very real harm, by assembling chains of inferences connecting individuals to harmful facts.[134]

b. Examples of Bad Anonymization

Several people have expressed the opinion that the three stories I describe highlight only the peril of bad anonymization.[135] These people have argued that the State of Massachusetts, AOL, and Netflix should have foreseen the vulnerability of their approaches to anonymization.[136] I have many responses.

First, and most fundamentally, the phrase "bad anonymization" is redundant. At least for forget-and-release methods, computer scientists have documented theoretical limits about the type of privacy that can be achieved, which I describe below.[137] Although some researchers have developed new techniques that do better than forget-and-release anonymization, these techniques have significant limitations, and I explore both the techniques and limitations below.[138]

Second, the fact that such sophisticated data handlers were responsible for these three data releases belies the idea that these were the mistakes of amateurs. Indeed, Netflix boasted about how it perturbed the Netflix Prize data before it released it to protect privacy.[139] Likewise, AOL's data release was stewarded by PhDs who seemed aware that they were dealing with sensitive information and approved by high-ranking officials.[140] With hindsight it is easy to argue that these breaches were foreseeable—nobody questions anymore

134. *See infra* Part III.A (describing "the database of ruin").

135. *E.g.*, Khaled El Emam, *Has There Been a Failure of Anonymization?*, ELECTRONIC HEALTH INFORMATION & PRIVACY, Aug. 19, 2009, http://ehip.blogs.com/ehip/2009/08/has-there-been-a-failure-of-anonymization.html ("Ohm has taken examples of poorly de-identified datasets that were re-identified and drew broad conclusions from those.").

136. *Id.*

137. *See infra* Part III.B.1.

138. *See infra* Part III.B.2 and III.B.3.

139. Netflix Prize: FAQ, *supra* note 92 ("Even if, for example, you knew all your own ratings and their dates you probably couldn't identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and that data was subject to perturbation.").

140. Zeller, Jr., *supra* note 78.

whether search queries can be used to identify users—but the past failure of foresight by sophisticated data handlers should give us pause about present claims of bad anonymization.

Third, when one considers the mistakes that have been made by sophisticated data handlers, one can begin to imagine the mistakes being made by the legions of less-sophisticated data handlers, the thousands of IT professionals with no special training in anonymization who are responsible for anonymizing millions of corporate databases. Even if we can divide anonymization cases into good and bad piles, it is safe to assume that the bad towers over the good.

Finally, even if we could teach every data handler in the world how to avoid the mistakes of the past—a daunting and expensive proposition—our new, responsible approach to anonymization would still do nothing to protect all of the data anonymized in the past. Database owners could reanonymize databases they still controlled, but they would not be able to secure the data they shared or redistributed in the past.

c. The Problem of Public Release

It would also be a mistake to conclude that the three stories demonstrate only the peril of public release of anonymized data. Some might argue that had the State of Massachusetts, AOL and Netflix kept their anonymized data to themselves, or at least shared the data much less widely, we would not have had to worry about data privacy.

There is obviously some logic to this objection. In Part IV, I argue that regulators should treat publicly released data differently than privately used data.[141]

On the other hand, we should not be surprised that we learned the lessons of reidentification only after public releases of data. Reidentification researchers can only reidentify that which they can access. But other people with access to less-public information might be reidentifying in private, keeping the results to themselves. Any time data is shared between two private parties, we should worry about the possibility of reidentification.

Moreover, we must not forget that anonymization is also used by companies as an internal privacy control—to allow Department A to share data with Department B without breaching customer privacy.[142] Just because data is kept wholly within a company does not put to rest concerns about expectations

141. *Infra* Part IV.C.1.
142. *See supra* notes 16–17 and accompanying text.

of privacy. If a company promises, for example, to share behavioral data with its marketing arm only in anonymized form, we should worry that the power of easy reidentification gives the company the tools needed to break that promise.

d. The Myth of the Superuser

Finally, some might object that the fact that reidentification is possible does not necessarily make it likely to happen. In particular, if there are no motivated, skilled adversaries, then there is no threat. I am particularly sensitive to this objection, because I have criticized those who try to influence policy by exploiting fears of great power, a tactic that relies on what I have called the "Myth of the Superuser."[143]

The power of reidentification, however, is not a Myth of the Superuser story for three reasons: First, reidentification techniques are not Superuser techniques. The Netflix study reveals that it is startlingly easy to reidentify people in anonymized data.[144] 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.[145] Second, the AOL release reminds us about the power of a small group of bored bloggers. And third, there are great financial motivations pushing people to reidentify.[146]

Moreover, I did not claim that feats of great power never happen online. Such a conclusion is provably false. Instead, I argued that because it is so easy to exaggerate power, we should hold those offering stories about online power to try to influence policy to a high standard of proof.[147] I concede that my claim of reidentification power should be held to the high standard of proof, and I argue that I have met that standard.

143. *See generally* Paul Ohm, *The Myth of the Superuser: Fear, Risk, and Harm Online*, 41 U.C. DAVIS L. REV. 1327 (2008).

144. *Netflix Prize Study*, *supra* note 5, at 112.

145. The INNER JOIN command is taught in beginner database texts. *See, e.g.*, OPPEL & SHELDON, *supra* note 131; TAYLOR, *supra* note 131, at 309; WILSON & COLBY, *supra* note 131, at 501.

146. *See* Salvador Ochoa et al., *Reidentification of Individuals in Chicago's Homicide Database: A Technical Legal Study* (unpublished student paper) (2001), *available at* http://web.mit.edu/sem083/www/assignments/reidentification.html (discussing financial motives pressing people to reidentify including those affecting marketers and blackmailers).

147. *See* Ohm, *supra* note 143, at 1402.

4. The Intuition Gap

What each of the foregoing objections highlights is the gap in intuition that persists among privacy experts. These privacy experts, primarily lawyers and business executives charged with protecting their companies' users, clients, and customers, cling to the idea that although anonymization may be weaker than we assumed, it has not failed. They may concede the need to change privacy policies or invest a bit more heavily in technology and expertise in response to the studies cited above, but they hope they need only small tweaks like these and not overhauls.

In the meantime, I predict that computer scientists and talented amateurs will continue to release new examples of powerful reidentification, with each announcement shaking those who still cling to false faiths. As have the past announcements, these future announcements will surprise experts by how cheaply, quickly, and easily supposedly robust anonymization will fall. I make these predictions confidently, because the power of reidentification traces two curves, both moving upward incessantly: the power of computer hardware and the richness of outside information.

The future of anonymization and reidentification thus promises years of awkward transition, as the privacy experts on the wrong side of the intuition gap weaken and then finally abandon their faith in anonymization. It may take years—maybe five, maybe more—before most privacy experts accept that they should abandon faith in anonymization, and these will be years filled with dashed hopes and recalibrated expectations. The gap will probably take longer to close than it fairly should, as companies and other interests vested in the cheap, easy promises of anonymization will try to convince others to persist in their faith despite the evidence.

The rest of this Article will mostly skip past the coming, painful years of transition while the intuition gap closes. Instead, it will plan for what happens next, after the intuition gap closes, once we realize that anonymization has failed. What does the failure of anonymization mean for privacy law?

II. HOW THE FAILURE OF ANONYMIZATION DISRUPTS PRIVACY LAW

Policymakers cannot simply ignore easy reidentification, because for decades they enacted laws and regulations while laboring under the robust anonymization assumption. They must now reexamine every privacy law and regulation to see if the easy reidentification result has thwarted their original designs.

Modern privacy laws tend to act preventatively, squeezing down the flow of particular kinds of information in order to reduce predictable risks of harm. In order to squeeze but not cut off valuable transfers of information, legislators have long relied on robust anonymization to deliver the best of both worlds: the benefits of information flow and strong assurances of privacy. The failure of anonymization has exposed this reliance as misguided, throwing carefully balanced statutes out of equilibrium.

At the very least, legislators must abandon the idea that we protect privacy when we do nothing more than identify and remove PII. The idea that we can single out fields of information that are more linkable to identity than others has lost its scientific basis and must be abandoned.

A. The Evolution of Privacy Law

In the past century, the regulation of privacy in the United States and Europe has evolved from scholarly discussion, to limited common law torts, to broad statutory schemes. Before deciding how to respond to the rise of easy reidentification, we must recognize three themes from this history of privacy law. First, while privacy torts focus solely on compensating injured victims of privacy harms, more recent privacy statutes shift the focus from post hoc redress to problem prevention. Second, this shift has led to the hunt for PII through quasi-scientific exercises in information categorization. Third, legislatures have tried to inject balance into privacy statutes, often by relying on robust anonymization.

1. The Privacy Torts: Compensation for Harm

Most legal scholars point to a celebrated nineteenth-century law review article by Samuel Warren and Louis Brandeis, *The Right to Privacy*,[148] as the wellspring of information privacy law. In the article, Warren and Brandeis, alarmed by the rise of tabloid journalism, advocated a new right of privacy, urging courts to allow plaintiffs to bring new privacy torts.[149] The concept of harm—intangible, incorporeal harm to mere feelings, but harm all the same—loomed large in the article. For example, Warren and Brandeis describe victims of privacy deprivations as experiencing "mental suffering,"[150] "mental pain and distress, far greater than could be inflicted by mere bodily

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

149. Irwin R. Kramer, *The Birth of Privacy Law: A Century Since Warren and Brandeis*, 39 CATH. U. L. REV. 703, 709 (1990).

150. Warren & Brandeis, *supra* note 148, at 213.

injury,"[151] and "injury to the feelings."[152] That the authors focused on harm is unsurprising because the entire article is a call for "[a]n action of tort for damages in all cases."[153]

Seventy years later, William Prosser synthesized the case law inspired by Warren and Brandeis into the four privacy torts commonly recognized in U.S. jurisdictions today: (1) intrusion upon the plaintiff's seclusion or solitude, or into his private affairs, (2) public disclosure of embarrassing private facts about the plaintiff, (3) publicity that places the plaintiff in a false light in the public eye, and (4) appropriation, for the defendant's advantage, of the plaintiff's name or likeness.[154] All four require actual injury, as do all torts.[155]

2. Shift to Broad Statutory Privacy: From Harm to Prevention and PII

Courts took the lead during the evolution of the privacy torts,[156] while legislatures stayed mostly in the background, doing little more than occasionally codifying privacy torts.[157] Then, about forty years ago, legislatures began to move to the forefront of privacy regulation, enacting sweeping new statutory privacy protections. The fear of computerization motivated this shift.

In the 1960s, the U.S. government began computerizing records about its citizens, combining this data into massive databases. These actions sparked great privacy concerns.[158] Throughout the decade, commentators described threats to privacy from computerization and helped defeat several government proposals.[159] Spurred by this, in 1973 an advisory committee created by the secretary of health, education, and welfare issued a report that proposed a new framework called "Fair Information Principles" (FIPS).[160] The FIPS have

151. *Id.* at 196.
152. *Id.* at 197.
153. *Id.* at 219.
154. William L. Prosser, *Privacy*, 48 CAL. L. REV. 383 (1960). Prosser was also the reporter for the second Restatement of Torts, in which he also promulgated his four privacy torts. RESTATEMENT (SECOND) OF TORTS § 652B (1977).
155. W. PAGE KEETON ET AL., PROSSER & KEETON ON TORTS 5 (5th ed. 1984) (defining torts as "a body of law which is directed toward the compensation of individuals . . . for losses which they have suffered").
156. Prosser, *supra* note 154, at 386–89.
157. *E.g.*, N.Y. CIV. RIGHTS LAW §§ 50–51 (McKinney 2007).
158. PRISCILLA M. REGAN, LEGISLATING PRIVACY: TECHNOLOGY, SOCIAL VALUES, AND PUBLIC POLICY 82 (1995).
159. Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 506–07 & nn.138–45 (2006).
160. U.S. DEP'T OF HEALTH, EDUC., & WELFARE, RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS (1973).

been enormously influential, inspiring statutes,[161] law review articles,[162] and multiple refinements.[163]

FIPS require a data protection scheme that provides, among other things, notice and consent, access, data integrity, enforcement, and remedies,[164] but for the present discussion, what the FIPS say is less important than what the FIPS wrought: a very different approach to privacy law, one that embraces rights of privacy that do more than solely redress past harm. Influenced by the FIPS, legislatures have enacted statutes designed to avoid "privacy problems" that have nothing to do with the "injury to feelings" at the heart of the privacy torts. As Dan Solove puts it, "These problems are more structural in nature. . . . They involve less the overt insult or reputational harm to a person and more the creation of the risk that a person might be harmed in the future."[165]

Thus, beginning in the 1970s, Congress began to enact statutes designed to reduce the risk of harm. Congress's approach for crafting these laws is best described as Linnaean. After first identifying a problem—"a risk that a person might be harmed in the future"[166]—lawmakers try to enumerate and categorize types of information that contribute to the risk. They categorize on a macro level (distinguishing between health information, education information, and financial information) and on a micro level (distinguishing between names, account numbers, and other specific data fields). Through this process, they have filled many pages of the U.S. Code with taxonomies of information types that deserve special treatment because of their unusual tendency to cause harm.[167]

Congress has thus embraced a wholly data-centric approach, the PII approach, to protecting privacy. This approach assumes that lawmakers can evaluate the inherent riskiness of data categories, assessing with mathematical precision whether or not a particular data field contributes to the problem enough to be regulated. In doing so, it tends to ignore messier, human factors

161. *E.g.*, The Privacy Act of 1974 "requires agencies to follow the Fair Information Practices when gathering and handling personal data." Daniel J. Solove & Chris Jay Hoofnagle, *A Model Regime of Privacy Protection*, 2006 U. ILL. L. REV. 357, 361 (citing 5 U.S.C. § 552a(e) (2000)).

162. *E.g.*, Marc Rotenberg, *Fair Information Practices and the Architecture of Privacy (What Larry Doesn't Get)*, 2001 STAN. TECH. L. REV. 1; Paul M. Schwartz, *Preemption and Privacy*, 118 YALE L.J. 902, 906–22 pt. I (2009).

163. ORGANISATION FOR ECONOMIC COOPERATION & DEV., OECD GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA (2001), *available at* http://www.uhoh.org/oecd-privacy-personal-data.PDF; Federal Trade Commission, *Fair Information Practice Principles*, http://www.ftc.gov/reports/privacy3/fairinfo.shtm (last visited June 12, 2010).

164. Federal Trade Commission, *supra* note 163.

165. Solove, *supra* note 159, at 487–88.

166. *Id.*

167. *See infra* notes 203–207 (giving examples of statutes that list categories of information).

that should also factor into a risk assessment, such as the likelihood that someone will be motivated enough to care about a particular dataset.[168]

It is necessary, however, to distinguish between two very different legislative motivations for singling out categories of information. The easy reidentification result calls into question only the second of these motivations. First, some statutes restrict sensitive information, the kind of information that causes fully-realized harm when disclosed.[169] For example, the Driver's Privacy Protection Act (DPPA) singles out "highly restricted personal information," including sensitive categories like "photograph" and "medical or disability information."[170] Easy reidentification has not disrupted the logic of provisions like this one. Even though robust anonymization has failed, it still makes sense to treat specially those kinds of information that can be used directly to cause harm.

In contrast, lawmakers often single out categories of data for special treatment under the mistaken belief that these categories (and only these) increase the linkability of anonymized data. For instance, the DPPA singles out a second category of personal information, including linkable data fields like social security number and driver identification number, for special, but less restrictive, treatment.[171] The law implicitly assumes that this list includes every data field that can link database records to identity—but easy reidentification proves otherwise. When legislators focus on linkability and identifiability in this way, they enshrine release-and-forget, deidentification, PII-removal approaches to anonymization into law. This approach to legislation makes little sense in light of the advances in easy reidentification.

3. How Legislatures Have Used Anonymization to Balance Interests

Writing about the privacy torts, William Prosser said that "[i]n determining where to draw the line the courts have been invited to exercise nothing less than a power of censorship over what the public may be permitted to read."[172] So too is every privacy statute an "exercise [in] the power of censorship."[173] These laws restrict the free flow of information. This should give lawmakers

168. *See infra* Part IV.B (discussing motive).

169. Arvind Narayanan & Vitaly Shmatikov, *De-Anonymizing Social Networks*, http://userweb.cs.utexas.edu/~shmat/shmat_oak09.pdf, app. B (last visited June 12, 2010) (noting that some laws single out information that "itself is sensitive," while others seek to prevent "deductive disclosure"). This paper was later published without appendices. *See De-Anonymizing Social Networks*, *supra* note 117.

170. 18 U.S.C. § 2725(3)-(4) (2006).

171. *Id.*

172. Prosser, *supra* note 154, at 413.

173. *Id.*

great pause. The free flow of information fuels the modern economy, nourishes our hunger for knowledge, shines a light on the inner workings of powerful institutions and organizations, and represents an exercise of liberty.[174] Before enacting any privacy law, lawmakers should weigh the benefits of unfettered information flow against its costs and must calibrate new laws to impose burdens only when they outweigh the harms the laws help avoid.

But for the past forty years, legislators have deployed a perfect, silver bullet solution—anonymization—that has absolved them of the need to engage in overt balancing. Anonymization liberated lawmakers by letting them gloss over the measuring and weighing of countervailing values like security, innovation, and the free flow of information. Regardless of whether those countervailing values weighed heavily, moderately, or barely at all, they would always outweigh the minimized risk to privacy of sharing anonymized data, which lawmakers believed to be almost nil thanks to anonymization. The demise of robust anonymization will throw the statutes legislatures have written out of balance, and lawmakers will need to find a new way to regain balance lost.

Consider how legislatures in two jurisdictions have relied upon anonymization to bring supposed balance to privacy law: the U.S.'s Health Insurance Portability and Accountability Act (HIPAA) and the EU's Data Protection Directive.

a. How HIPAA Used Anonymization to Balance Health Privacy

In 1996, the U.S. Congress enacted the Health Insurance Portability and Accountability Act (HIPAA), hoping to improve healthcare and health insurance in this country.[175] Among the other things it accomplishes, HIPAA is a significant privacy law. Title II of the Act mandates compliance with health privacy regulations, which have been promulgated by the Department

174. *See* Kent Walker, *Where Everybody Knows Your Name: A Pragmatic Look at the Costs of Privacy and the Benefits of Information Exchange*, 2000 STAN. TECH. L. REV. 2, 7–21 (enumerating the benefits of shared information).

175. Pub. L. No. 104-191, 110 Stat. 1936 (1996). According to the preamble to the Act, the purpose of HIPAA is:

> To amend the Internal Revenue Code of 1986 to improve portability and continuity of health insurance coverage in the group and individual markets, to combat waste, fraud, and abuse in health insurance and health care delivery, to promote the use of medical savings accounts, to improve access to long-term care services and coverage, to simplify the administration of health insurance, and for other purposes.

Id.

of Health and Human Services (HHS) and are now known as the HIPAA Privacy Rule.[176]

In many ways, the HIPAA Privacy Rule represents the high-water mark for use of PII to balance privacy risks against valuable uses of information.[177] HIPAA demonstrates Congress's early sensitivity to the power of reidentification, through its treatment of what it calls the "de-identification of health information" (DHI).[178] HIPAA itself exempts data protected by DHI from any regulation whatsoever,[179] but defines DHI so as to allow for further regulatory interpretation—and HHS has used this statutory mandate to define DHI as information that "does not identify an individual" nor provide "a reasonable basis to believe that the information can be used to identify an individual."[180]

HHS's Privacy Rule elaborates this vague reasonability standard further in two alternate ways. First, under the so-called "statistical standard," data is DHI if a statistician or other "person with appropriate knowledge . . . and experience" formally determines that the data is not individually identifiable.[181] Second, data is DHI under the so-called "safe harbor standard" if the covered entity suppresses or generalizes eighteen enumerated identifiers.[182] The Privacy Rule's list is seemingly exhaustive—perhaps the longest such list in any privacy regulation in the world. Owing to the release of Dr. Sweeney's study around the same time, the Privacy Rule requires the researcher to generalize birth dates to years[183] and ZIP codes to their initial three digits.[184]

Congress and HHS concluded simply that by making data unidentifiable, health professionals could trade sensitive information without impinging on patient privacy. Moreover, they froze these conclusions in amber, enumerating a single, static list, one they concluded would protect privacy in all health privacy contexts.[185] In promulgating the Privacy Rule, regulators relied on their

176. *Id.* § 264 (directing the secretary of Health and Human Services to submit standards for protecting privacy); HIPAA Privacy Rule, 45 C.F.R. §§ 160, 164 (2009).

177. Jay Cline, *Privacy Matters: When Is Personal Data Truly De-Identified?*, COMPUTERWORLD, July 24, 2009, http://www.computerworld.com/s/article/9135898/Privacy_matters_When_is_personal_data_truly_de_identified ("No other country has developed a more rigorous or detailed guidance for how to convert personal data covered by privacy regulations into non-personal data."). HIPAA is not the most recent information privacy law enacted in the U.S. *See, e.g.*, Gramm-Leach-Bliley Act of 1999, Pub. L. No. 106-102, (codified as 15 U.S.C. §§ 6801–6809 (2006)); Children's Online Privacy Protection Act of 1998, Pub. L. No. 106-170, (codified as 15 U.S.C. §§ 6501–6506 (2006)).

178. *See* 45 C.F.R. §§ 164.502(d)(2), 164.514(a), (b) (2009).

179. *Id.*

180. *Id.* § 164.514(a).

181. *Id.* § 164.514(b)(1).

182. *Id.* § 164.514(b)(2).

183. *Id.* § 164.514(b)(2)(C).

184. *Id.* § 164.514(b)(2)(B) (allowing only two digits for ZIP codes with 20,000 or fewer residents).

185. Since promulgating the safe harbor list almost a decade ago, HHS has never amended it.

faith in the power of anonymization as a stand-in for a meaningful cost-benefit balancing. This is an opportunity lost, because it is hard to imagine another privacy problem with such starkly presented benefits and costs. On one hand, free exchange of information among medical researchers can help them develop treatments to ease human suffering and save lives. On the other hand, medical secrets are among the most sensitive we hold. It would have been quite instructive to see regulators explicitly weigh such stark choices.

By enumerating eighteen identifiers, the Privacy Rule assumes that any other information that might be contained in a health record cannot be used to reidentify. We now understand the flaw in this reasoning, and we should consider revising the Privacy Rule as a result.[186]

b. How the EU Data Protection Directive Used Anonymization to Balance Internet Privacy

EU lawmakers have also relied upon the power of anonymization to avoid difficult balancing questions. Unlike the American approach with HIPAA, however, the EU enacted a broad, industry-spanning law,[187] the Data Protection Directive, which purports to cover any "personal data" held by any data administrator.[188] Data is personal data if it can be used to identify someone "directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."[189]

The EU never intended the Directive to apply to all data. Instead, it meant for "personal data" to exclude at least some data—data that was not "directly or indirectly" identifiable, such as anonymized data—from regulation. Like their U.S. counterparts, EU lawmakers imagined they could strike a balance through the power of technology. If anonymization worked, data administrators could freely share information so long as data subjects were no longer "directly or indirectly" identifiable. With this provision, EU lawmakers sought to preserve space in society for the storage and transfer of anonymized data, thereby providing room for unencumbered innovation and free expression.

186. *See infra* Part IV.D.1.

187. The Directive obligates EU countries to transpose its rules into domestic laws within a set time frame. Eur. Comm'n Justice & Home Affairs, *Transposition of the Data Protection Directive*, http://ec.europa.eu/justice_home/fsj/privacy/lawreport/index_en.htm (last visited June 12, 2010).

188. EU Data Protection Directive, *supra* note 3, art. 2(a).

189. *Id.*

Whether and to what extent the Directive retains such a preserve has been debated in the internet privacy context.[190] For several years, the EU has clashed with companies like Google, Yahoo, and Microsoft over what they must do to protect databases that track what their users do online.[191] Much of this debate has turned on what companies must do with stored IP addresses. An IP address is a numeric identifier assigned to every computer on the internet.[192] Just as a social security number identifies a person, an IP address identifies a computer, so an IP address can tie online conduct to location and identity.[193] Every computer reveals its IP address to every other computer it contacts,[194] so every time I visit Google, my computer reveals its IP address to a Google computer.[195] Following longstanding industry practice, Google records my IP address along with details about what I am doing when using Google's services.[196]

Google has argued to the EU that it protects the privacy of its users using anonymization, by throwing away part, not all, of every IP address.[197] Specifically, an IP address is composed of four equal pieces called *octets*,[198] and Google stores the first three octets and deletes the last, claiming that this practice protects user privacy sufficiently.[199] Google's competitors, Microsoft and Yahoo, are much more thorough, throwing away entire IP addresses.[200]

At its core, this too is a debate about balance—between the wonderful innovations Google promises it can deliver by studying our behavior,[201] and the

190. *See, e.g.*, Frederick Lah, Note, *Are IP Addresses "Personally Identifiable Information?"*, 4 I/S: J.L. & POL'Y FOR INFO. SOC'Y 681 (2008).

191. *E.g.*, Posting of Saul Hansell, *Europe: Your IP Address Is Personal*, N.Y. TIMES BITS BLOG, http://bits.blogs.nytimes.com/2008/01/22/europe-your-ip-address-is-personal (Jan. 22, 2008).

192. DOUGLAS COMER, 1 INTERNETWORKING WITH TCP/IP 42 (5th ed. 2006).

193. *Id.* at 43–44.

194. *Id.* at 35–36.

195. *Id.*

196. SIMSON GARFINKEL & GENE SPAFFORD, WEB SECURITY, PRIVACY AND COMMERCE 211 (2002).

197. Letter From Google to Congressman Joe Barton 14–15 (Dec. 21, 2007), *available at* http://searchengineland.com/pdfs/071222-barton.pdf.

198. COMER, *supra* note 192, at 53.

199. Letter From Google to Congressman Joe Barton, *supra* note 197, at 14–15.

200. *Behavioral Advertising: Industry Practice and Consumers' Expectations, Hearings Before the H. Comm. on Energy and Commerce, Subcomm. on Communications, Technology and the Internet and Subcomm. on Commerce, Trade and Consumer Protection*, 111th Cong. 1 (2009) (statement of Anne Toth, Head of Privacy, Yahoo! Inc.); Posting of Peter Cullen, Chief Privacy Strategist at Microsoft, Microsoft Privacy & Safety, *Microsoft Supports Strong Industry Search Data Anonymization Standards*, MICROSOFT PRIVACY AND SAFETY BLOG, http://blogs.technet.com/privacyimperative/archive/2008/12/08/microsoft-supports-strong-industry-search-data-anonymization-standards.aspx (Dec. 8, 2008).

201. In 2008, to try to placate those worried about privacy, Google authored a series of blog posts "about how [they] harness the data [they] collect to improve [their] products and services for [their] users." *E.g.*, Posting of Matt Cutts, Software Engineer, *Using Data to Fight Webspam*, THE OFFICIAL

possible harm to users whose IP addresses are known or revealed. Again, claims that we should trust robust anonymization stand in for nuanced, careful cost-benefit balancing arguments. Google promises we can have our cake while it eats it too—by placing our trust in data anonymization.

B. How the Failure of Anonymization Disrupts Privacy Law

In addition to HIPAA and the EU Data Protection Directive, almost every single privacy statute and regulation[202] ever written in the U.S. and the EU embraces—implicitly or explicitly, pervasively or only incidentally—the assumption that anonymization protects privacy, most often by extending safe harbors from penalty to those who anonymize their data. At the very least, regulators must reexamine every single privacy law and regulation. The loss of robust anonymization reveals the lurking imbalance in these privacy laws, sometimes shifting in favor of protecting privacy too much and sometimes favoring the flow of information too much.

Easy reidentification makes PII-focused laws like HIPAA underprotective by exposing the arbitrariness of their intricate categorization and line drawing. Although HIPAA treats eighteen categories of information as especially identifying,[203] it excludes from this list data about patient visits—like hospital name, diagnosis, year of visit, patient's age, and the first three digits of ZIP code—that an adversary with rich outside information can use to defeat anonymity.

Many other laws follow the same categorization-and-line-drawing approach. The Driver's Privacy Protection Act requires special handling for "personal information" including, among other things, "social security number, driver identification number, name, address . . . , [and] telephone number,"[204] while requiring much less protection of "the 5-digit zip code" and "information on vehicular accidents, driving violations, and driver's status."[205] Similarly, the Federal Education Rights and Privacy Act (FERPA) singles out for protection "directory information," including, among other things, "name,

---

GOOGLE BLOG, http://googleblog.blogspot.com/2008/06/using-data-to-fight-webspam.html (June 27, 2008, 4:51 EST) (linking to earlier posts in the series).

202. In this Article, I focus on statutes and regulations for several reasons. First, these rules provide a concrete set of texts about which I can make correspondingly concrete observations. Second, American and European approaches to privacy legislation differ somewhat, providing a comparative study. Third, when it comes to dictating how information is collected, analyzed, and disclosed in modern life, no other source of law has the influence of privacy statutes and regulations.

203. 45 C.F.R. §§ 164.502(d)(2), 164.514(a), (b) (2009).

204. 18 U.S.C. § 2725(3) (2006).

205. *Id.*

address, telephone listing, date and place of birth, [and] major field of study."[206] Federal Drug Administration regulations permit the disclosure of "records about an individual" associated with clinical trials "[w]here the names and other identifying information are first deleted."[207] These are only a few of many laws that draw lines and make distinctions based on the linkability of information. When viewed in light of the easy reidentification result, these provisions, like HIPAA, seem arbitrary and underprotective.

In contrast, easy reidentification makes laws like the EU Data Protection Directive overbroad—in fact, essentially boundless. Because the Directive turns on whether information is "directly or indirectly" linked to a person,[208] each successful reidentification of a supposedly anonymized database extends the regulation to cover that database. As reidentification science advances, it expands the EU Directive like an ideal gas to fit the shape of its container. A law that was meant to have limits is rendered limitless, disrupting the careful legislative balance between privacy and information and extending data-handling requirements to all data in all situations.

Notice that the way the easy reidentification result disrupts the Directive is the mirror image of the way it impacts HIPAA. Easy reidentification makes the protections of HIPAA illusory and underinclusive because it deregulates the handling of types of data that can still be used to reidentify and harm. On the other hand, easy reidentification makes laws like the EU Data Protection Directive boundless and overbroad. We should tolerate neither result because both fail to achieve the balance that was originally at the heart of both types of laws.

Most privacy laws match one of these two forms. Even the few that do not fit neatly into one category or the other often contain terms that are made indeterminate and unpredictable by easy reidentification. As one example, the Stored Communications Act in the U.S. applies to "record[s] or other information pertaining to a subscriber . . . or customer," without specifying what degree of identifiability makes a record "pertain."[209] As reidentification science advances, courts will struggle to decide whether anonymized records fall within this definition. The vagueness of provisions like this will invite costly litigation and may result in irrational distinctions between jurisdictions and between laws.

206. 20 U.S.C. § 1232g(a)(5)(A) (2006).
207. 21 C.F.R. § 21.70(a)(3)(i) (2009).
208. EU Data Protection Directive, *supra* note 3, art. 2(a).
209. 18 U.S.C. § 2702(c) (2006).

C. The End of PII

1. Quitting the PII Whack-a-Mole Game

At the very least, we must abandon the pervasively held idea that we can protect privacy by simply removing personally identifiable information (PII). This is now a discredited approach. Even if we continue to follow it in marginal, special cases, we must chart a new course in general.

The trouble is that PII is an ever-expanding category. Ten years ago, almost nobody would have categorized movie ratings and search queries as PII, and as a result, no law or regulation did either.[210] Today, four years after computer scientists exposed the power of these categories of data to identify, no law or regulation yet treats them as PII.

Maybe four years has not been enough time to give regulators the chance to react. After all, HIPAA's Privacy Rule, which took effect in 2003, does incorporate Dr. Sweeney's research, conducted in the mid-1990s.[211] It expressly recognizes the identifying power of ZIP code, birth date, and sex, and carves out special treatment for those who delete or modify them, along with fifteen other categories of information.[212] Should this be the model of future privacy law reform—whenever reidentification science finds fields of data with identifying power, should we update our regulations to encompass the new fields? No. This would miss the point entirely.

HIPAA's approach to privacy is like the carnival whack-a-mole game: As soon as you whack one mole, another will pop right up. No matter how effectively regulators follow the latest reidentification research, folding newly identified data fields into new laws and regulations, researchers will always find more data field types they have not yet covered.[213] The list of potential PII will never stop growing until it includes everything.[214]

Consider another reidentification study by Narayanan and Shmatikov.[215] The researchers have reidentified anonymized users of an online social network based almost solely on the stripped-down graph of connections between

---

210. The Video Privacy Protection Act, enacted in 1988, protects lists of movies watched not because they are PII, but because they are sensitive. 18 U.S.C. § 2710 (2006). For more on the distinction, see *supra* Part II.A.2.

211. *See supra* Part I.B.1.b (describing Sweeney's research).

212. 45 C.F.R. §§ 164.502(d)(2), 164.514(a)-(b) (2009).

213. *See* Narayanan & Shmatikov, *supra* note 169 ("While some data elements may be uniquely identifying on their own, *any* element can be identifying in combination with others.").

214. *Cf. id.*; Dinur & Nissim, *supra* note 115, at 202 ("[T]here usually exist other means of identifying patients, via indirectly identifying attributes stored in the database.").

215. *See* Narayanan & Shmatikov, *supra* note 169.

people.[216] By comparing the structure of this graph to the nonanonymized graph of a different social network, they could reidentify many people even ignoring almost all usernames, activity information, photos, and every other single piece of identifying information.[217]

To prove the power of the method, the researchers obtained and anonymized the entire Twitter social graph, reducing it to nameless, identity-free nodes representing people connected to other nodes representing Twitter's "follow" relationships. Next, they compared this mostly deidentified husk of a graph[218] to public data harvested from the Flickr photo-sharing social-network site. As it happens, tens of thousands of Twitter users are also Flickr users, and the researchers used similarities in the structures of Flickr's "contact" graph and Twitter's "follow" graph to reidentify many of the anonymized Twitter user identities. With this technique, they could reidentify the usernames or full names of one-third of the people who subscribed to both Twitter and Flickr.[219] Given this result, should we add deidentified husks of social networking graphs—a category of information that is almost certainly unregulated under U.S. law, yet shared quite often[220]—to the HIPAA Privacy Rule list and to the lists in other PII-focused laws and regulations? Of course not.

Instead, lawmakers and regulators should reevaluate any law or regulation that draws distinctions based solely on whether particular data types can be linked to identity, and should avoid drafting new laws or rules grounded in such a distinction. This is an admittedly disruptive prescription. PII has long served as the center of mass around which the data privacy debate has orbited.[221] But although disruptive, this proposal is also necessary. Too often, the only thing that gives us comfort about current data practices is that an administrator has gone through the motions of identifying and deleting PII—and in such cases, we deserve no comfort at all. Rather, from now on we need a new organizing principle, one that refuses to play the PII whack-a-mole game. Anonymization has become "privacy theater";[222] it should no longer be considered to provide meaningful guarantees of privacy.

---

216. *See De-Anonymizing Social Networks*, *supra* note 117, at 182–85.

217. *Id.* at 184.

218. *Id.* To make their study work, the researchers first had to "seed" their data by identifying 150 people who were users of both Twitter and Flickr. They argue that it would not be very difficult for an adversary to find this much information, and they explain how they can use "opportunistic seeding" to reduce the amount of seed data needed. *Id.* at 181–85.

219. *Id.*

220. *Id.* at 174–75 (surveying examples of how social-network data is shared).

221. *See* Leslie Ann Reis, *Personally Identifiable Information*, *in* 2 ENCYCLOPEDIA OF PRIVACY 383–85 (William G. Staples ed., 2006).

222. Paul M. Schwartz, *Reviving Telecommunications Surveillance Law*, 75 U. CHI. L. REV. 287, 310–15 (2008) (developing the concept of privacy theater).

2. Abandoning "Anonymize" and "Deidentify"

We must also correct the rhetoric we use in information privacy debates. We are using the wrong terms, and we need to stop. We must abolish the word anonymize;[223] let us simply strike it from our debates. A word that should mean, "try to achieve anonymity" is too often understood to mean "achieve anonymity," among technologists and nontechnologists alike. We need a word that conjures effort, not achievement.

Latanya Sweeney has similarly argued against using forms of the word "anonymous" when they are not literally true.[224] Dr. Sweeney instead uses "deidentify" in her research. As she defines it, "[i]n deidentified data, all explicit identifiers, such as SSN, name, address, and telephone number, are removed, generalized, or replaced with a made-up alternative."[225] Owing to her influence, the HIPAA Privacy Rule explicitly refers to the "de-identification of protected health information."[226]

Although "deidentify" carries less connotative baggage than "anonymize," which might make it less likely to confuse, I still find it confusing. "Deidentify" describes release-and-forget anonymization, the kind called seriously into question by advances in reidentification research. Despite this, many treat claims of deidentification as promises of robustness,[227] while in reality, people can deidentify robustly or weakly.[228] Whenever a person uses the unmodified word "deidentified," we should demand details and elaboration.

Better yet, we need a new word for privacy-motivated data manipulation that connotes only effort, not success. I propose "scrub." Unlike "anonymize" or "deidentify," it conjures only effort. One can scrub a little, a lot, not enough,

223. Anonymize is a relatively young word. The Oxford English Dictionary traces the first use of the word "anonymized" to 1972 by Sir Alan Marre, the UK's Parliamentary Ombudsman. OXFORD ENGLISH DICTIONARY (Additions Series 1997) ("I now lay before Parliament . . . the full but anonymised texts of . . . reports on individual cases."). According to the OED, the usage of the word is "chiefly for statistical purposes." *Id.*

224. Latanya Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, 25 J.L. MED. & ETHICS 98, 100 (1997) ("The term *anonymous* implies that the data cannot be manipulated or linked to identify an individual.").

225. *Id.*

226. 45 C.F.R. § 164.514(a) (2009) (defining term).

227. *See, e.g.*, *infra* Part IV.D.2.a (discussing Google's weak approach to anonymization of search engine log files and how the company treats these practices as robust).

228. For similar reasons, I do not recommend replacing "anonymize" with the parallel construction "pseudonymize." *See* Christopher Soghoian, *The Problem of Anonymous Vanity Searches*, 3 I/S: J.L. & POL'Y FOR INFO SOC'Y 299, 300 (2007) ("In an effort to protect user privacy, the records were 'pseudonymized' by replacing each individual customer's account I.D. and computer network address with unique random numbers."). Just as "anonymize" fails to acknowledge reversible scrubbing, "pseudonymize" fails to credit robust scrubbing.

or too much, and when we hear the word, we are not predisposed toward any one choice from the list. Even better, technologists have been using the word scrub for many years.[229] In fact, Dr. Sweeney herself has created a system she calls Scrub for "locating and replacing personally-identifying information in medical records."[230]

III. HALF MEASURES AND FALSE STARTS

Focusing on things other than PII is a disruptive and necessary first step, but it is not enough alone to restore the balance between privacy and utility that we once enjoyed. How do we fix the dozens, perhaps hundreds, of laws and regulations that we once believed reflected a finely calibrated balance, but in reality rested on a fundamental misunderstanding of science? Before turning, in Part IV, to a new test for restoring the balance lost, let us first consider three solutions that are less disruptive to the status quo but are unfortunately also less likely to restore the balance. Legislators must understand why these three solutions—which they will be tempted to treat as the only necessary responses—are not nearly enough, even in combination, to restore balance to privacy law.

First, lawmakers might be tempted to abandon the preventative move of the past forty years, taking the failure of anonymization as a signal to return to a regime that just compensates harm. Even if such a solution involves an aggressive expansion of harm compensation—with new laws defining new types of harms and increasing resources for enforcement—this is a half measure, a necessary but not sufficient solution. Second, lawmakers might be encouraged to wait for the technologists to save us. Unfortunately, although technologists *will* develop better privacy-protection techniques, they will run up against important theoretical limits. Nothing they devise will share the single-bullet universal power once promised by anonymization, and thus any technical solutions they offer must be backed by regulatory approaches. Finally, some will recommend doing little more than banning reidentification. Such a ban will almost certainly fail.

---

229. *See, e.g.*, Jeremy Kirk, *Yahoo to Scrub Personal Data After Three Months*, IDG NEWS SERVICE, Dec. 17, 2008, *available at* http://www.pcworld.com/article/155610/yahoo_to_scrub_personal_data_after_three_months.html (reporting Yahoo!'s decision to "anonymize" its databases of sensitive information ninety days after collection); Tommy Peterson, *Data Scrubbing*, COMPUTERWORLD, Feb. 10, 2003, http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=78230.

230. Latanya Sweeney, *Replacing Personally-Identifying Information in Medical Records, the Scrub System*, *in* 1996 J. AM. MED. INFORMATICS ASS'N PROC. 333.

A. Strictly Punish Those Who Harm

If reidentification makes it easier for malevolent actors like identity thieves, blackmailers, and unscrupulous advertisers to cause harm, perhaps we need to step up enforcement of preexisting laws prohibiting identity theft,[231] extortion,[232] and unfair marketing practices.[233] Anything we do to deter those who harm and provide remedies for those harmed is, in light of the increased power of reidentification, imperative. But this is merely a necessary response, not a sufficient one.

Full retreat to a tort-based privacy regime, which would abandon the forty-year preventative turn in privacy law, would be a grave mistake, because without regulation, the easy reidentification result will spark a frightening and unprecedented wave of privacy harm by increasing access to what I call the "database of ruin." The database of ruin exists only in potential: It is the worldwide collection of all of the facts held by third parties that can be used to cause privacy-related harm to almost every member of society. Easy access to the database of ruin flows from what I call the "accretion problem."

1. The Accretion Problem

The accretion problem is this: Once an adversary has linked two anonymized databases together, he can add the newly linked data to his collection of outside information and use it to help unlock other anonymized databases. Success breeds further success. Narayanan and Shmatikov explain that "once any piece of data has been linked to a person's *real* identity, any association between this data and a *virtual* identity breaks the anonymity of the latter."[234] This is why we should worry even about reidentification events that seem to expose only nonsensitive information, because they increase the *linkability* of data, and thereby expose people to potential future harm.

Because of the accretion problem, every reidentification event, no matter how seemingly benign, brings people closer to harm. Recall that Narayanan and Shmatikov linked two IMDb users to records in the Netflix Prize database. To some online observers, this connection seemed nonthreatening and trivial[235] because they did not care if others knew what movies they had rented.

---

231. *E.g.*, 18 U.S.C. § 1028 (2000); CAL. PENAL CODE § 530.5 (1999); MASS. GEN. LAWS ch. 266, § 37E (2002); N.Y. PENAL LAW §§ 190.77–190.84 (2010).

232. *E.g.*, 18 U.S.C. § 872 (2006) (prohibiting extortion by federal government officials).

233. *E.g.*, 15 U.S.C. § 45 (2006) (FTC provision regulating unfair competition); CAL. BUS. & PROF. CODE §§ 17200–210 (2008).

234. *Netflix Prize Study*, *supra* note 5, at 119.

235. *E.g.*, Comment of chef-ele to Netflix Prize Forum, http://www.netflixprize.com/community/

These people failed to see how connecting IMDb data to Netflix data is a step on the path to significant harm. Had Narayanan and Shmatikov not been restricted by academic ethical standards (not to mention moral compunction), they might have connected people to harm themselves.

The researchers could have treated the connections they made between IMDb usernames and Netflix Prize data as the middle links in chains of inferences spreading in two directions: one toward living, breathing people and the other toward harmful facts. For example, they could have tied the list of movies rated in the Netflix Prize database to a list of movies rated by users on Facebook. I suspect that the fingerprint-like uniqueness of Netflix movie preferences would hold for Facebook movie preferences as well.[236]

They could have also easily extended the chain in the other direction by making one reasonable assumption: People tend to reuse usernames at different websites.[237] User john_doe20 on IMDb is likely to be john_doe20 on many other websites as well.[238] Relying on this assumption, the researchers could have linked each living, breathing person revealed through Facebook, through the Netflix Prize data, through IMDb username, to a pseudonymous user at another website. They might have done this with noble intentions. Perhaps they could have unearthed the identity of the person who had savagely harassed people on a message board.[239] Maybe they could have determined who had helped plan an attack on a computer system on a 4chan message board.[240] But they also could have revealed identities to evil ends. Perhaps they could have tied identities to the pseudonymous people chatting on a child abuse victims' support website, in order to blackmail, frame, or embarrass them.

---

viewtopic.php?id=809 (Nov. 28, 2007, 09:04:54) ("I think you can find out more about a person by typing their name into Google; this Netflix data reverse-engineering doesn't seem to be a bigger threat than that."); Comment of jimmyjot to The Physics arXiv Blog, http://arxivblog.com/?p=142, (Feb. 17, 2008) ("Choice of movies also does not tell a whole lot."). See also various comments to the posting *Anonymity of Netflix Prize Dataset Broken*, SLASHDOT, http://it.slashdot.org/article.pl?sid=07/11/27/1334244&from=rss (Nov. 27, 2007).

236. Of course, even without the Netflix data release, Narayanan and Shmatikov might have been able to connect some records in the IMDb database directly to Facebook records. But recall that for many users, the Netflix data contains movies not rated in IMDb. I am assuming that for some of the people who use all three services, no direct connection between IMDb and Facebook is possible. Thanks to Jane Yakowitz for this point.

237. Arvind Narayanan, *Lendingclub.com: A De-Anonymization Walkthrough*, 33 BITS OF ENTROPY BLOG, http://33bits.org/2008/11/12/57 (Nov. 12, 2008) ("Many people use a unique username everywhere . . . ."); *De-Anonymizing Social Networks*, *supra* note 117, at 6–7 (relying on fact that users tend to reuse usernames on different social networks).

238. *See* Narayanan, *supra* note 237.

239. Danielle Keats Citron, *Cyber-Civil Rights*, 89 B.U. L. REV. 61, 71–75 (2009) (discussing harassing comments on the AutoAdmit internet discussion board).

240. Mattathias Schwartz, *The Trolls Among Us*, N.Y. TIMES MAG., Aug. 3, 2008, at MM24 (describing 4chan).

Imagine a large-scale attack on the pseudonyms used on the social networking site Experience Project, which tries to connect users to people who have had similar life experiences.[241] If the researchers had access to other, harder-to-obtain, outside information, they could have caused even greater harm. With access to Google's search query log file, they might have learned the diseases people had been recently looking up.[242] By connecting the IMDb usernames to Facebook biographies, they might have been able to bypass password recovery mechanisms for their victims' online email and bank accounts, allowing them to steal private communications or embezzle money, just as somebody broke into Sarah Palin's email account by guessing that she had met her husband at "Wasilla high."[243] Other possible mischief is easy to imagine when one considers databases that track criminal histories, tax payments, bankruptcies, sensitive health secrets like HIV status and mental health diagnoses, and more.

2. The Database of Ruin

It is as if reidentification and the accretion problem join the data from all of the databases in the world together into one, giant, database-in-the-sky, an irresistible target for the malevolent. Regulators should care about the threat of harm from reidentification because this database-in-the-sky contains information about all of us.

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.

---

241. Experience Project, About Us, http://www.experienceproject.com/about.php (last visited July 5, 2010).

242. *See infra* Part IV.D.2.b (discussing the risk to privacy from access to search query logs).

243. *See* Posting of Sam Gustin, *Alleged Palin Email Hacker Explains*, PORTFOLIO.COM TECH OBSERVER BLOG, http://www.portfolio.com/views/blogs/the-tech-observer/2008/09/18/alleged-palin-email-hacker-explains (Sept. 18, 2008).

3. Entropy: Measuring Inchoate Harm

But even regulators who worry about the database of ruin will probably find it hard to care about the reidentification of people to nonsensitive facts like movie ratings. Until there is completed harm—until the database of ruin is accessed—they will think there is no need to regulate. One way to understand the flaw in this is through the concept of entropy.[244]

In thermodynamics, entropy measures disorder in a system; in information theory, it tracks the amount of information needed to describe possible outcomes.[245] Similarly, in reidentification science, entropy measures how close an adversary is to connecting a given fact to a given individual.[246] It describes the length of the inference chains heading in opposite directions, quantifying the remaining uncertainty.

Consider entropy in the children's game, Twenty Questions.[247] At the start of a game, the Answerer thinks of a subject the Questioner must discover through yes or no questions. Before any questions have been asked, entropy sits at its maximum because the Answerer can be thinking of any subject in the world. With each question, entropy decreases, as each answer eliminates possibilities. The item is a vegetable; it is smaller than a breadbox; it is not green. The Questioner is like the reidentifier, connecting outside information to the anonymized database, reducing entropic uncertainty about the identity of his target.

Entropy formalizes the accretion problem. We should worry about reidentification attacks that fall short of connecting anonymized data to actual identities, and we should worry about reidentification attacks that do not reveal sensitive information. Even learning a little benign information about a supposedly anonymized target reduces entropy and brings an evil adversary closer to his prey.

Consider one more extended metaphor, which Part IV builds upon to illustrate a prescription.[248] Imagine each person alive stands on one side of a long hallway specifically dedicated just for him or her. At the other end of the hallway sits that person's ruinous fact, the secret their adversary could use to cause them great harm. In the hallway between the person and the ruinous

244. Arvind Narayanan, *About 33 Bits*, 33 BITS OF ENTROPY BLOG, http://33bits.org/about (Sept. 28, 2008) (explaining the concept of entropy).

245. The concept originated with a seminal paper by Claude Shannon. *See* C.E. Shannon, *A Mathematical Theory of Communication*, 27 BELL SYS. TECH. J. 379 (1948).

246. Narayanan, *supra* note 244.

247. I am indebted to Anna Karion for the analogy.

248. *See infra* Part IV.A.

fact, imagine a long series of closed, locked doors, each lock requiring a different key, which represent the database fields that must be reconnected or the links in the inferential chain that must be established to connect the person to the fact. Finally, imagine many other people clutching keys to some of the doors. Each person represents a database owner, and the keys the person holds represent the inferences the person can make, using the data they own.

Under the current, now discredited PII approach to privacy regulation, we tend to hold database owners—the people in the middle of the hallway—accountable for protecting privacy only if they happen to hold one of two critical keys. First, if they hold the key that unlocks the first door, the one closest to the data subject, we regulate them. This is the linkability form of PII.[249] Second, if they hold the key that unlocks the last door, the one closest to the ruinous fact, we also regulate them. This is the sensitivity form of PII.[250] But under our current approach, we tend to immunize all of the database owners whose keys unlock only doors in the middle of the hallway.

4. The Need to Regulate Before Completed Harm

If we fail to regulate reidentification that has not yet ripened into harm, then adversaries can nudge each of us ever closer to the brink of connection to our personal database of ruin. It will take some time before most people become precariously compromised, and whether it will take months, years, or decades is difficult to predict. Because some people have more to hide than others, the burden of decreasing entropy will not be distributed equally across society.[251]

Once we are finally connected to our databases of ruin, we will be unable to unring the bell. As soon as Narayanan and Shmatikov tied an IMDb username to Netflix rental data, they created an inferential link in the chain, and no regulator can break that link. Anybody who wants to can replicate their result by downloading the Netflix Prize data[252] and mining the IMDb

249. *See supra* notes 169–171 and accompanying text (explaining difference between sensitive and linkable forms of PII).

250. *See id*.

251. There are two classes of people who may escape this fate altogether: those with no secrets and those so disconnected from the grid that databases hold few records about them—including many residents of lesser-developed countries. In our own advanced society, I tend to believe that the numbers of people in these groups are so small that they are like myths—the unicorns and mermaids of information privacy. Ultimately, the size of these groups is a difficult empirical question, but one that is not particularly important. I think most people would agree that large majorities in advanced societies are susceptible to reidentification harms, making privacy regulation an important question for huge parts of the world.

252. Since the competition is now over, the data is no longer publicly available, but it has already been downloaded hundreds of times. *Netflix Prize Study*, *supra* note 5, at 119.

user ratings database. Narayanan and Shmatikov have forever reduced the privacy of the people whose information they connected. The FBI cannot easily order connected databases unconnected, nor can they confiscate every last copy of a particularly harmful database.

If we worry about the entire population being dragged irreversibly to the brink of harm, we must regulate in advance because hoping to regulate after the fact is the same as not regulating at all. So long as our identity is separated from the database of ruin by a high degree of entropy, we can rest easy. But as data is connected to data, and as adversaries whittle down entropy, every one of us will soon be thrust to the brink of ruin.

B. Wait for Technology to Save Us

Regulators may wonder whether the technologists will save us first. If we view parallel advances in reidentification and anonymization as an arms race, even though the reidentifiers have raced ahead for now, perhaps the anonymizers will regain the advantage through some future breakthrough. Maybe such a breakthrough will even restore the status quo and shift the privacy laws back into balance.

We should not expect a major breakthrough for release-and-forget anonymization, because computer scientists have proved theoretical limits of the power of such techniques. The utility and privacy of data are linked, and so long as data is useful, even in the slightest, then it is also potentially reidentifiable. Moreover, for many leading release-and-forget techniques, the tradeoff is not proportional: As the utility of data increases even a little, the privacy plummets.

We might, however, enjoy some help from new technology, although we should not expect a breakthrough. Computer scientists have devised techniques that are much more resistant to reidentification than release-and-forget. Data administrators may use some of these techniques—interactive techniques, aggregation, access controls, and audit trails—to share their data with a reduced risk of reidentification. Alas, despite the promise of these techniques, they cannot match the sweeping privacy promises that once were made regarding release-and-forget anonymization. The improved techniques tend to be much slower, more complex, and more expensive than simple anonymization. Worse, these techniques are useless for many types of data analysis problems. Technological advances like these may provide some relief in a post-anonymization, post-PII world, but they can never replace the need for a regulatory response.

1. Why Not to Expect a Major Breakthrough

Computer scientists have begun to conclude that in the arms race between release-and-forget anonymization and reidentification, the reidentifiers hold the permanent upper hand.

a. Utility and Privacy: Two Concepts at War

Utility and privacy are, at bottom, two goals at war with one another.[253] In order to be useful, anonymized data must be imperfectly anonymous. "[P]erfect privacy can be achieved by publishing nothing at all—but this has no utility; perfect utility can be obtained by publishing the data exactly as received from the respondents, but this offers no privacy."[254] No matter what the data administrator does to anonymize the data, an adversary with the right outside information can use the data's residual utility to reveal other information. Thus, at least for useful databases, perfect anonymization is impossible.[255] Theorists call this the impossibility result.[256] There is always some piece of outside information that could be combined with anonymized data to reveal private information about an individual.[257]

Cynthia Dwork offers proof of the impossibility result.[258] Although useful data can never be perfectly private, it is important to understand the practical limits of this result;[259] some kinds of theoretical privacy breach may concern policymakers very little. To use Dwork's example, if a database owner releases an aggregate statistic listing the average heights of women in the world by national origin, an adversary who happens to know that his target is precisely two inches shorter than the average Lithuanian woman may learn a "private" fact by studying the database.[260] Although we would properly say that the utility of the anonymized data revealed a private fact when combined with outside information,[261] we would be foolhardy to regulate or forbid the release of databases containing aggregated height data to avoid this possibility. In

253. Shuchi Chawla et al., *Toward Privacy in Public Databases*, *in* 2 THEORY CRYPTOGRAPHY CONF. 363 (2005).
254. *Id.* at 364.
255. Dwork, *supra* note 122, at 4.
256. *Id.*
257. Dinur & Nissim, *supra* note 115, at 203 (showing, for a particular model, "tight impossibility results," meaning that privacy would require "totally ruining the database usability").
258. Dwork, *supra* note 122.
259. *Id.*
260. *Id.*
261. *Id.*

this case, the richness of the outside information creates almost all of the privacy breach, and the statistic itself contributes very little.

Although the impossibility result should inform regulation, it does not translate directly into a prescription. It does not lead, for example, to the conclusion that all anonymization techniques are fatally flawed, but instead, as Cynthia Dwork puts, "to a new approach to formulating privacy's goals."[262] She calls her preferred goal "differential privacy" and ties it to so-called interactive techniques. Differential privacy and interactive techniques are discussed below.

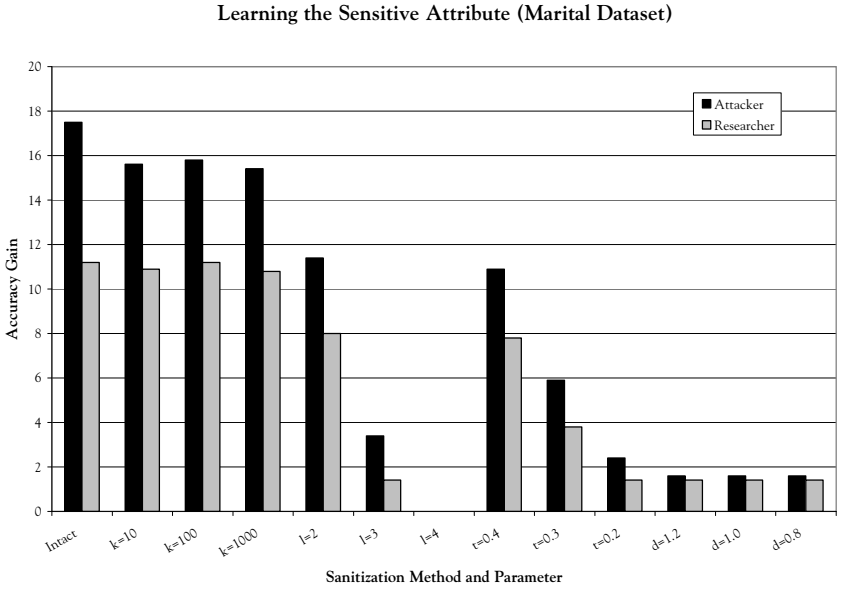
b. The Inverse and Imbalanced Relationship

Other theoretical work suggests that release-and-forget anonymization techniques are particularly ill-suited for protecting privacy while preserving the utility of data. Professor Shmatikov, one of the Netflix Prize researchers, coauthored a study with Justin Brickell that offers some depressing insights about the tradeoffs between utility and privacy for such techniques. As the researchers put it, "even modest privacy gains require almost complete destruction of the data-mining utility."[263]

The researchers compared several widely used anonymization techniques to a form of anonymization so extreme no data administrator would ever use it: a completely wiped database with absolutely no information beyond the single field of information under study[264]—for a health study perhaps the diagnoses, for an education study the grade point averages, and for a labor study the salaries. We would hope that real-world anonymization would compare very favorably to such an extreme method of anonymization, of course supplying worse privacy, but in exchange preserving much better utility.[265] Although the full details are beyond the scope of this Article, consider the intuition revealed in the following graph:

262. *Id.*
263. Brickell & Shmatikov, *supra* note 48, at 70, 76.
264. *Id.* at 70–71.
265. *See id.*

FIGURE 1: Effects on Privacy and Utility of Anonymization[266]



In Figure 1, the pairs of bars represent the same database transformed into many different forms using widespread anonymization techniques. For each pair, the left, black bar represents the privacy of the data, with smaller bars signifying more privacy. The right, gray bars represent the utility of the data, with longer bars meaning more utility. Anonymization techniques search for ways to shorten the left bar without shortening the right bar too much, and the holy grail of anonymization would be a short, black bar next to a long, gray bar. Even a quick scan of the graph reveals the absence of this condition.

The leftmost pair of bars, with a privacy score of almost eighteen and a utility score of about eleven, represents the original, unadulterated data. A score of zero represents the utility or privacy of completely wiped data. Notice how the first three pairs of bars, the ones labeled with the letter k, describe techniques that preserve a lot of utility while improving privacy very little.[268]

266. This figure has been adapted from a figure in *id.* at 76. Only the formatting has been changed; the substance of the figure remains the same.

268. These bars represent techniques that achieve *k-anonymity*, a widely embraced metric for strong anonymity. *Id.* at 71; Sweeney, *supra* note 8 (defining *k-anonymity*).

Although the second trio of bars, those labeled with the letter l,[269] show much greater improvements in privacy than the first trio, such improvements come only at great losses to utility.

These results show that for traditional, widespread, release-and-forget anonymization, not only are privacy and utility related, but their relationship is skewed. Small increases in utility are matched by even bigger decreases in privacy, and small increases in privacy cause large decreases in utility. The researchers concluded that even the most sophisticated anonymization techniques were scarcely better than simply throwing away almost all of the data instead.

Thus, using traditional, release-and-forget, PII-focused anonymization techniques, any data that is even minutely useful can never be perfectly anonymous, and small gains in utility result in greater losses for privacy. Both of these relationships cut against faith in anonymization and in favor of other forms of regulation.

2. The Prospect of Something Better Than Release-and-Forget

Researchers have developed a few techniques that protect privacy much better than the traditional, release-and-forget techniques. These work by relaxing either the release or the forget requirement. For example, some data administrators never release raw data, releasing only aggregated statistics instead. Every day, *USA Today* summarizes a survey in a colorful graph on their front page. Armed only with these survey responses, it would be very difficult for a reidentifier to prove that any particular person took part in a *USA Today* survey, much less gave a particular response.

Similarly, some researchers favor interactive techniques.[270] With these techniques, the data administrator answers questions about the data without ever releasing the underlying data. For example, an analyst might ask, what percentage of the people in your database have been diagnosed with this rare form of cancer? This might prompt the administrator to calculate and return the answer—say, 2 percent. In most cases, reidentifiers will find it much more difficult to link answers like these to identity than if they had access to the underlying raw data.

---

269. These bars represent *l-diversity*, another widely adopted metric. The final six bars represent *t-closeness*. Brickell & Shmatikov, *supra* note 48, at 70–71.

270. Cynthia Dwork et al., *Calibrating Noise to Sensitivity in Private Data Analysis*, *in* 2006 THEORY CRYPTOGRAPHY CONF. 265, 267.

Researchers can do even better. Using one class of interactive techniques, those that satisfy a requirement called differential privacy,[271] the data administrator never even releases the accurate statistic; instead, she introduces a carefully calculated amount of random noise to the answer, ensuring mathematically that even the most sophisticated reidentifier will not be able to use the answer to unearth information about the people in the database.[272]

Finally, just as these techniques refer to something less than full release, other techniques refuse to forget—instead, they monitor what happens to data *after* release. Borrowing from computer security research, these techniques involve the use of access controls and audit trails.[273] Using these techniques, data administrators release their data but only after protecting it using software that limits access and tracks usage. The data analyst who receives the protected data will be able to interact with it only in limited ways, and the analyst's every move will be recorded in the audit trail and reported back to the data administrator or a third-party watchdog.

3. The Limitations of the Improved Techniques

Unfortunately, these alternatives do not make up for the broken promises of release-and-forget anonymization. For starters, they tend to be less flexible than traditional anonymization. Interactive techniques require constant participation from the data administrator. This increases the cost of analysis and reduces the rate of new analysis. Because an analyst must submit requests and wait for responses, he is not free to simply test theory after theory at the maximum rate. Even worse, without access to the raw data, he might miss useful research inquiries that reveal themselves to those who study trends in the data.

Furthermore, even with interactive techniques and aggregation, data administrators cannot promise perfect privacy. As an example, if an adversary somehow knows that his target is the only man who visited a hospital clinic Thursday afternoon, then the aggregated answer to the question, "diagnoses of men who visited the clinic Thursday afternoon" reveals sensitive information tied directly to an identity. As another example, despite decades of denials from the Census Bureau, scholars have unearthed proof that the agency provided aggregated, city-block-level data that helped locate Japanese Americans who were then sent to internment camps during the Second World

271. *See* Dwork, *supra* note 122, at 8–9.

272. *See* Adam & Wortmann, *supra* note 60, at 540 (describing the "output-perturbation approach").

273. For more information on access controls in the computer security context, see RICK LEHTINEN ET AL., COMPUTER SECURITY BASICS 66–72 (2006).

War.[274] Even though the data did not identify particular houses or families, just telling authorities how many Japanese lived on each block gave them enough information to do enormous harm.

Interactive techniques that introduce noise are also of limited usefulness. For example, a city may want to analyze census data to determine where to run a bus line to serve elderly residents. Noise introduced to provide privacy may inadvertently produce the wrong answer to this question.[275] Similarly, law enforcement data miners may find it unacceptable to tell a judge that they are using a "noisy" technique to justify asking for a search warrant to search a home.[276] Techniques that satisfy differential privacy also require complex calculations that can be costly to perform.[277]

Finally, computer security researchers have thoroughly documented the problem with creating robust access controls.[278] Simply put, even the best computer security solutions are bug-prone, as well as being expensive to create and deploy.[279] All of these reasons explain why the vast majority of data shared or stored today is protected—if at all—by traditional, release-and-forget anonymization, not by these more exotic, more cumbersome, and more expensive alternatives.

Even if computer scientists tomorrow develop a groundbreaking technique that secures data much more robustly than anything done today—and this is a very unlikely "if"—the new technique will only work on data secured in the future; it will do nothing to protect data that has been stored or disclosed in the past. A database, once released, can become easier to reidentify but never more difficult. Long chains of inferences from past reidentification cannot be broken with tomorrow's advances.

Techniques that eschew release-and-forget may improve over time, but because of inherent limitations like those described above, they will never supply

274. William Seltzer & Margo Anderson, Population Association of America, After Pearl Harbor: The Proper Role of Population Data Systems in Time of War (Mar. 28, 2000) (unpublished paper), *available at* https://pantherfile.uwm.edu/margo/www/govstat/newpaa.pdf.

275. *See* Chawla et al., *supra* note 253, at 366.

276. The difficulty of using "noisy" techniques in police work is illustrated by a recent AP story that documents one instance where the addition of "random material" to a database resulted in repeated unnecessary police deployments. *Cops: Computer Glitch Led to Wrong Address*, MSNBC NEWS, Mar. 19, 2010, http://www.msnbc.msn.com/id/35950730.

277. Jon Kleinberg et al., *Auditing Boolean Attributes*, *in* 2000 ACM SYMP. ON PRINCIPLES DATABASE SYS. 86 (proving that particular method supporting interactive technique is NP-hard, meaning computationally expensive).

278. BRUCE SCHNEIER, BEYOND FEAR: THINKING SENSIBLY ABOUT SECURITY IN AN UNCERTAIN WORLD 87–101 (2003).

279. *Id.*; *cf.* FREDERICK P. BROOKS, JR., THE MYTHICAL MAN-MONTH (1975) (discussing how software engineering principles lead to bugs).

a silver-bullet alternative. Technology cannot save the day, and regulation must play a role.

C. Ban Reidentification

Finally, some have urged simply banning reidentification.[280] Lawmakers can offer a straightforward argument for a ban: By anonymizing data, a data administrator gives notice of her intent to protect the privacy of her data subjects, who may rely on this notice when consenting to provide her their data. A reidentifying adversary thwarts this intent and undermines this consent so much that we might need a law banning the act itself.

A reidentification ban is sure to fail, however, because it is impossible to enforce. How do you detect an act of reidentification?[281] Reidentification can happen completely in the shadows. Imagine that Amazon.com anonymizes its customer purchase database and transmits it to a marketing firm. Imagine further that although the marketing firm promises not to reidentify people in Amazon's database, it could increase profits significantly by doing so. If the marketing firm breaks its promise and reidentifies, how will Amazon or anybody else ever know? The marketing firm can conduct the reidentification in secret, and gains in revenue may not be detectable to the vendor.

This problem appears insurmountable, although four forces might help to ameliorate it. First, lawmakers might pair a ban with stricter penalties and better enforcement, for example by declaring reidentification a felony and providing extra money to the FBI and FTC for enforcement. Second, lawmakers can give citizens a private right of action against those who reidentify.[282] Third, lawmakers can mandate software audit trails for those who use anonymized data.[283] Finally, a smaller scale ban, one imposed only on trusted recipients of specific databases—for example, a ban prohibiting government data-miners from reidentifying—may be much easier to enforce.[284]

---

280. Earl Lane, *A Question of Identity: Computer-Based Pinpointing of 'Anonymous' Health Records Prompts Calls for Tighter Security*, NEWSDAY, Nov. 21, 2000, at C8 (quoting Janlori Goldman, head of the Health Privacy Project at Georgetown University as saying: "Our goal has been to get a national policy making it illegal to re-identify an anonymized database").

281. *Id.* ("As long as the data recipient is discreet, an agency may never learn if its information is being compromised." (citing Latanya Sweeney)).

282. They can model this on the Federal Stored Communications Act, which provides a civil cause of action to any "person aggrieved by any violation" of the Act. 18 U.S.C. § 2707 (2006).

283. *E.g.*, 45 C.F.R. § 164.308(a)(1)(ii)(D) (2009) (describing HIPAA Security Rule mandating "Information system activity review" including regular review of "audit logs").

284. For another example, see *infra* Part IV.D.1 (discussing the ban on reidentification for trusted recipients of health information).

I predict that any of these marginal improvements would still be outweighed by the inherent difficulty of detecting secret reidentification for private gain. This significant detection problem makes a ban extremely unlikely to succeed.

IV. RESTORING BALANCE TO PRIVACY LAW AFTER THE FAILURE OF ANONYMIZATION

Once regulators conclude that the three partial solutions discussed above are not enough to restore balance to privacy law after the failure of anonymization, they must do more. They should weigh the benefits of unfettered information flow against the costs of privacy harms. They should incorporate risk assessment strategies that deal with the reality of easy reidentification as the old PII model never could. Ultimately, they should consider a series of factors to identify situations in which harm is likely and whether it outweighs the benefits of unfettered information flow. When they identify harm that outweighs these benefits, they should regulate, focusing on narrow contexts and specific sectors rather than trying to regulate broadly across industries. To demonstrate how this approach works, this Part ends with two case studies recommending new strategies for regulating the privacy of health and internet usage information.

A. Which Database Owners Should We Regulate Anew?

In the search for a new organizing principle to supplement PII, I start from the premise that any privacy rule we devise must distinguish between different types of database owners and different types of databases. This approach might sound like PII, but it is broader. The problem is not that the PII approach categorizes; the problem is that it focuses on only a few, narrowly drawn categories that seem insufficient and even somewhat arbitrary in light of easy reidentification. Recall the hallway metaphor: PII-based rules regulate only those people with a key to the first door closest to the data subject (those that can link to a user's identity) or a key to the last door closest to the ruinous fact (those holding sensitive information).[285] For example, HIPAA singles out for special treatment social security numbers (linkable data) and medical diagnoses (sensitive data). PII rules ignore the people who can unlock doors only in the middle.

285. *See* text accompanying *supra* notes 248–250.

The power of reidentification demands that we begin to regulate the middle. But how? It would be logically justifiable but overly aggressive to regulate any entity possessing any fragment of data at any point along the chain of inferences, covering even a person holding only one key. We should aim to direct scarce regulatory resources at those database owners that most contribute to the risk of the database of ruin through well-tuned rules. A rule that regulates the database owner in the middle that possesses but a single scrap of unimportant data puts too much regulatory focus on too slight a risk.

Which database owners in the middle most contribute to the risk of harm and thereby most deserve government scrutiny and regulation? To the current PII-approach—regulation for those holding linkable data and those holding sensitive data—I propose we add at least one more category of database owners, the "large entropy reducers."[286] Large entropy reducers are entities that amass massive databases containing so many links between so many disparate kinds of information that they represent a significant part of the database of ruin, even if they delete from their databases all particularly sensitive and directly linkable information.

We can justify treating these entities differently using the language of duty and fault. Because large entropy reducers serve as one-stop shops for adversaries trying to link people to ruinous facts, they owe their data subjects a heightened duty of care. When a large entropy reducer loses control of its massive database, it causes much more harm than an entity holding much less data.

Who are large entropy reducers? In the hallway metaphor, they are the people clutching many keys; imagine the mythical janitor's keyring, jangling with dozens of different keys. In practice, this category includes large credit agencies like Experian, TransUnion, and Equifax; commercial data brokers like ChoicePoint, Acxiom, and LexisNexis; and internet search providers like Google, Microsoft, and Yahoo. These are among the most important large entropy reducers, but there are many others, and we should develop a more precise definition of the category, perhaps one taking advantage of the formal definition of entropy.

---

286. In addition to large entropy reducers, other classes of database owners probably deserve new regulation to account for the way they increase the risk of harm due to easy reidentification. For one, some database owners can make links between fields of information that can be connected by few other people—they can unlock doors requiring keys held by few people. For example, consider how a cell phone provider or automobile toll booth administrator can track physical movement and location in ways that few other providers can. Likewise, some database owners hold fields of data that act as identifiers on many sites, making them powerful tools for reidentification. Increasingly, email addresses act in this manner, as websites use them in place of usernames. Perhaps any entity holding an email address deserves new regulation. I plan in future work to develop these categories further, and to flesh out the arguments for regulating them more closely.

We should expand existing privacy laws and enact new privacy laws that regulate the behavior of companies like these. To be sure, many of these firms are already obligated to comply with many different privacy laws, but in light of easy reidentification and the database of ruin, we need to regulate them more, perhaps with new rules tailored to limiting the type of risk of reidentification such providers represent.

B. Regulatory Principles

Now that we know whom to regulate—database owners holding linkable or sensitive data (PII) and large entropy reducers—we turn to the content of regulation. How should regulators respond to the power of reidentification and the collapse of our faith in anonymization? Before we turn to a list of factors that will guide us to the proper regulation, we need to understand some overarching principles. This step is necessary because so much of how we regulate privacy depends on our faith in anonymization; stripped of this faith, we need to reevaluate some core principles.

1. From Math to Sociology

Regulators need to shift away from thinking about regulation, privacy, and risk only from the point of view of the data, asking whether a particular field of data viewed in a vacuum is identifiable. Instead, regulators must ask a broader set of questions that help reveal the risk of reidentification and threat of harm. They should ask, for example, what has the data administrator done to reduce the risk of reidentification? Who will try to invade the privacy of the people in the data, and are they likely to succeed? Do the history, practices, traditions, and structural features of the industry or sector instill particular confidence or doubt about the likelihood of privacy?

Notice that while the old approach centered almost entirely on technological questions—it was math and statistics all the way down—the new inquiry is cast also in sociological, psychological, and institutional terms. Because easy reidentification has taken away purely technological solutions that worked irrespective of these messier, human considerations, it follows that new solutions must explore, at least in part, the messiness.[287]

287. *See* Chawla et al., *supra* note 253, at 367 (noting that the relative advantage of one interactive technique is that "the real data can be deleted or locked in a vault, and so may be less vulnerable to bribery of the database administrator").

2. Support for Both Comprehensive and Contextual Regulation

The failure of anonymization will complicate one of the longest-running debates in information privacy law: Should regulators enact comprehensive, cross-industry privacy reform, or should they instead tailor specific regulations to specific sectors?[288] Usually, these competing choices are labeled, respectively, the European and United States approaches. In a postanonymization world, neither approach is sufficient alone: We need to focus on particular risks arising from specific sectors because it is difficult to balance interests comprehensively without relying on anonymization. On the other hand, we need a comprehensive regulation that sets a floor of privacy protection because anonymization permits easy access to the database of ruin. In aiming for both general and specific solutions, this recommendation echoes Dan Solove, who cautions that privacy should be addressed neither too specifically nor too generally.[289] Solove says that we should simultaneously "resolve privacy issues by looking to the specific context,"[290] while at the same time using "a general framework to identify privacy harms or problems and to understand why they are problematic."[291]

Thus, the U.S.'s exclusively sectoral approach is flawed, because it allows entire industries to escape privacy regulation completely based on the illusion that some data, harmless data, data in the middle of long chains of inferences leading to harm, is so bland and nonthreatening that it is not likely to lead to harm if it falls into the wrong hands. The principle of accretive reidentification shatters this illusion. Data almost always forms the middle link in chains of inferences, and any release of data brings us at least a little closer to our personal databases of ruin. For this reason, there is an urgent need for comprehensive privacy reform in this country. A law should mandate a minimum floor of safe data-handling practices on every data handler in the U.S. Further, it should require even stricter data-handling practices for every large entropy reducer in the U.S.

But on the other hand, the European approach—and specifically the approach the EU has taken in the Data Protection Directive—sets the height of this floor too high. Many observers have complained about the onerous

288. *See, e.g.*, Schwartz, *supra* note 162, at 908–16 (discussing history of sectoral and comprehensive approaches to privacy law).

289. DANIEL J. SOLOVE, UNDERSTANDING PRIVACY 46–49 (2008).

290. *Id.* at 48.

291. *Id.* at 49.

obligations of the Directive.[292] It might have made good sense to impose such strict requirements (notice, consent, disclosure, accountability) on data administrators when we still believed in the power of anonymization because the law left the administrators with a fair choice: Anonymize your data to escape these burdens or keep your data identifiable and comply.

But as we have seen, easy reidentification has mostly taken away this choice, thereby broadening the reach of the Directive considerably. Today, the EU hounds Google about IP addresses; tomorrow, it can make similar arguments about virtually any data-possessing company or industry. A European privacy regulator can reasonably argue that *a*n*y* database containing facts (no matter how well scrubbed) relating to people (no matter how indirectly) very likely now falls within the Directive. It can impose the obligations of the Directive even on those who maintain databases that contain nothing that a layperson would recognize as relating to an individual, so long as the data contains idiosyncratic facts about the lives of individuals.

I suspect that some of those who originally supported the Directive might feel differently about a Directive that essentially provides no exception for scrubbed data—a Directive covering most of the data in society. The Directive's aggressive data-handling obligations might have seemed to strike the proper balance between information flow and privacy when we thought that they were restricted to "personal data," but once reidentification science redefines "personal data" to include almost all data, the obligations of the Directive might seem too burdensome. For these reasons, the European Union might want to reconsider whether it should lower the floor of its comprehensive data-handling obligations.

Finally, once the U.S. tackles comprehensive privacy reform and the EU lowers the burdens of the directive, both governments should expand the process of imposing heightened privacy regulations on particular sectors. What might be needed above the comprehensive floor for health records may not be needed for phone records, and what might solve the problems of private data release probably will not work for public releases.[293] This approach borrows from Helen Nissenbaum, who urges us to understand privacy through what she calls "contextual integrity," which "couches its prescriptions always within the bounds of a given context" as better than other "universal"

---

292. *E.g.*, DOROTHEE HEISENBERG, NEGOTIATING PRIVACY: THE EUROPEAN UNION, THE UNITED STATES AND PERSONAL DATA PROTECTION 29, 30 (2005) (calling parts of the Directive "quite strict" and "overly complex and burdensome").

293. *Cf. infra* Part IV.D (discussing specific rules for health privacy and search engine privacy contexts).

accounts.[294] This approach also stands in stark contrast to the advice of other information privacy scholars and activists, who tend to valorize sweeping, society-wide approaches to protecting privacy and say nothing complimentary about the U.S.'s sectoral approach.

What easy reidentification thus demands is a combination of comprehensive data-protection regulation and targeted, enhanced obligations for specific sectors. Many others have laid out the persuasive case for a comprehensive data privacy law in the United States, so I refer the reader elsewhere for that topic.[295] The rest of the Article explores how to design sector-specific data privacy laws, now that we can no longer lean upon the crutch of robust anonymization to give us balance. What does a post-anonymization privacy law look like?

C. The Test

In the post-anonymization age, once regulators pick a target for regulation—say, large entropy reducers in the healthcare industry—they should weigh the following factors to determine the risk of reidentification in that context. The list is not exhaustive; other factors might be relevant.[296] The factors serve two purposes: They are indicators of risk and instruments for reducing risk. As indicators, they signal the likelihood of privacy harm. For example, when data administrators in a given context tend to store massive quantities of information, the risk of reidentification increases. Regulators should use these indicative factors like a score card, tallying up the risk of reidentification.

Once regulators decide to regulate, they should then treat these factors as instruments for reducing risk—the tuning knobs they can tweak through legislation and regulation to reduce the risk of harm. As only one example, regulators might ban public releases of a type of data outright while declining to regulate private uses of data.

---

294. Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 154 (2004).

295. *E.g.*, Solove & Hoofnagle, *supra* note 161.

296. The European privacy watchdog, the Article 29 Working Group, offers the following, similar but not identical, list of factors:

> The cost of conducting identification is one factor, but not the only one. The intended purpose, the way the processing is structured, the advantage expected by the controller, the interests at stake for the individuals, as well as the risk of organisational dysfunctions (e.g. breaches of confidentiality duties) and technical failures should all be taken into account. On the other hand [one] . . . should consider the state of the art in technology at the time of the processing and the possibilities for development during the period for which the data will be processed.

2007 Working Party Opinion, *supra* note 28, at 15.

1. Five Factors for Assessing the Risk of Privacy Harm

Data-Handling Techniques

How do different data-handling techniques affect the risks of reidentification? Experts probably cannot answer this question with mathematical precision; it is unlikely we can ever know, say, that the suppression of names and social security numbers produces an 82 percent risk, while interactive techniques satisfying differential privacy produce a 1 percent risk. Still, computer scientists could likely provide a rough relative ordering of different techniques—or at the very least, grade data-handling practices according to whether the risk of reidentification is high, medium, or low.[297] For example, computer scientists might grade favorably a database owner that uses the kind of new interactive techniques described earlier, although remember that such techniques are no panacea.

Private Versus Public Release

Regulators should scrutinize data releases to the general public much more closely than they do private releases between trusted parties. We fear the database of ruin because we worry that our worst enemy can access it, but if we use regulation to limit the flow of information to trusted relationships between private parties, we can breathe a little easier. It is no coincidence that every case study presented in Part I.B involved the public release of anonymized data. In each case, the researcher or researchers targeted the particular data because it was easy to get, and in the AOL search query example in particular, an army of blogger-reidentifiers acted as a force multiplier, aggravating greatly the breach and the harm.

My argument against public releases of data pushes back against a tide of theory and sentiment flowing in exactly the opposite direction. Commentators place great stock in the "wisdom of crowds," the idea that "all of us are smarter than any of us."[298] Companies like Netflix release great stores of information they once held closely to try to harness these masses.[299]

---

297. Some computer scientists have already tentatively offered studies that attempt to categorize the risk of reidentification of different techniques. *See, e.g.*, Lakshmanan et al., *supra* note 46 (focusing on anonymization); Adam & Wortmann, *supra* note 60 (evaluating methods, including conceptual, query restriction, data perturbation, and output perturbation). These studies do not take into account the latest advances in reidentification, but they are models for future work.

298. SUROWIECKI, *supra* note 15.

299. *See* Thompson, *supra* note 93.

The argument even throws some sand into the gears of the Obama Administration's tech-savvy new approach to governance. Through the launch of websites like data.gov[300] and the appointment of federal officials like CTO Aneesh Chopra[301] and CIO Vivek Kundra,[302] the administration has promised to release massive databases heralding a twenty-first century mode of government openness.[303] Amidst the accolades that have been showered upon the government for these efforts,[304] one should pause to consider the costs. We must remember that utility and privacy are two sides of the same coin,[305] and we should assume that the terabytes of useful data that will soon be released on government websites will come at a cost to privacy commensurate with, if not disproportionate to,[306] the increase in sunlight and utility.

Quantity

Most privacy laws regulate data quality but not quantity.[307] Laws dictate what data administrators can do with data according to the nature, sensitivity, and linkability of the information, but they tend to say nothing about how much data a data administrator may collect, nor how long the administrator can retain it. Yet, in every reidentification study cited, the researchers were aided by the size of the database. Would-be reidentifiers will find it easier to

300. Data.gov, About, http://www.data.gov/about (last visited June 12, 2010) ("The purpose of Data.gov is to increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government.").

301. *See* Posting of Nate Anderson, *Obama Appoints Virginia's Aneesh Chopra US CTO*, ARSTECHNICA LAW & DISORDER BLOG, http://arstechnica.com/tech-policy/news/2009/04/obama-appoints-virginias-aneesh-chopra-us-cto.ars (Apr. 20, 2009, 13:01 EST).

302. *See* Posting of Brian Knowlton, *White House Names First Chief Information Officer*, N.Y. TIMES CAUCUS BLOG, http://thecaucus.blogs.nytimes.com/2009/03/05/white-house-names-first-chief-information-officer (Mar. 5, 2009, 10:06 EST).

303. *Id.* ("Mr. Kundra discussed some of his plans and interests, including his intention . . . to create a data.gov web site that will put vast amounts of government information into the public domain.").

304. *E.g.*, Posting of Clay Johnson, *Redesigning the Government: Data.gov*, SUNLIGHTLABS.COM, http://www.sunlightlabs.com/blog/2009/04/16/redesigning-government-datagov (Apr. 16, 2009, 11:52 EST); Posting by Infosthetics, *Data.gov: How to Open Up Government Data*, INFORMATION AESTHETICS BLOG, http://infosthetics.com/archives/2009/03/open_up_government_data.html (Mar. 13, 2009, 17:25 EST). *But see* David Robinson, Harlan Yu, William P. Zeller & Edward W. Felten, *Government Data and the Invisible Hand*, 11 YALE J.L. & TECH. 160, 161 (2009) (discussing how the federal government should structure systems to enable greater internet-based transparency).

The Center for Democracy and Technology has posted a supportive but more cautious memo, flagging concerns about Data.gov involving deidentification and reidentification. Ctr. for Democracy & Tech., Government Information, Data.gov and Privacy Implications, http://www.cdt.org/policy/government-information-datagov-and-privacy-implications (July 13, 2009) ("While Data.gov has great potential, there are important privacy implications associated with data disclosure.").

305. *See supra* Part III.B.1.a.

306. *See supra* Part III.B.1.b.

307. *See supra* Part II.A.3 (listing privacy statutes that draw distinctions based on data type).

match data to outside information when they can access many records indicating the personal preferences and behaviors of many people. Thus, lawmakers should consider enacting new quantitative limits on data collection and retention.[308] They might consider laws, for example, mandating data destruction after a set period of time, or limiting the total quantity of data that may be possessed at any one time.

Motive

In many contexts, sensitive data is held only by a small number of actors who lack the motive to reidentify.[309] For example, rules governing what academic researchers can do with data should reflect the fact that academic researchers rarely desire to reidentify people in their datasets. A law that strictly limits information sharing for the general public—think FERPA (student privacy), HIPAA (health privacy), or ECPA (electronic communications privacy)—might be relaxed to allow researchers to analyze the data with fewer constraints. Of course, regulators should draw conclusions about motive carefully, because it is hard to predict who the adversary is likely to be, much less divine his or her motive.

Regulators should also weigh economic incentives for reidentification. Although we should worry about our enemies targeting us to learn about our medical diagnoses, we should worry even more about financially-motivated identity thieves looking for massive databases that they can use to target thousands simultaneously.[310]

Trust

The flip side of motive is trust. Regulators should try to craft mechanisms for instilling or building upon trust in people or institutions. While we labored

308. *See* European Union Article 29 Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Relating to Search Engines*, 00737/EN WP 148, at 19 (April 4, 2008), *available at* http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf [hereinafter 2008 Working Party Opinion] (arguing that search engines should store queries for a maximum of six months).

309. *Cf.* EU Data Protection Directive, *supra* note 3, at recital 26 (noting that "the means likely reasonably to be used" to identify individuals are relevant to a determination of whether individuals are "identifiable").

310. As one commentator puts it:

> [T]here's far less economic incentive for a criminal to go after medical data instead of credit card information. It's harder to monetize the fact that I know that Judy Smith of Peoria has heart disease—by filing false claims in her name, for example—than to have Judy's credit card number and expiration date. If I'm a criminal with advanced data skills and I have a day to spend, I'm going to go after financial data and not health data.

Cline, *supra* note 177.

under the shared hallucination of anonymization, we trusted the technology, so we did not have to trust the recipients of data; now that we have lost trust in the technology, we need to focus more on trust in people. We might, for example, conclude that we trust academic researchers implicitly, government data miners less, and third-party advertisers not at all, and we can build these conclusions into law and regulation.

2. Applying the Test

By applying the five factors, regulators will have a rough sense of the risk of reidentification of a particular type of provider in a particular context. If the risk is very low, regulators might choose to do nothing. If the risk is very high, regulators should feel inclined to act, imposing new restrictions on data collection, use, processing, or disclosure, and requiring specific data safe-handling procedures.

Regulators should perhaps also take into consideration the sensitivity of the data. It makes sense to treat medical diagnoses differently than television-watching habits, for example, because the path to harm for the former is shorter and more direct than for the latter. But because the database of ruin can be built almost entirely with nonsensitive data, regulators should beware not to make too much of this step in the analysis.

Finally, regulators should compare the risk and the sensitivity to the various benefits of unfettered information flow: for medical privacy, better treatments and saved lives; for internet privacy, better search tools and cheaper products; for financial privacy, fewer identity thefts. If the benefits of unfettered information significantly outweigh the costs to privacy in a particular context, they might decide to surrender.[311] Perhaps lawmakers will see reidentification as the latest example of the futility of attempting to foist privacy on an unappreciative citizenry through ham-handed regulations. Maybe they will conclude they should just give up and live in a society with very little privacy.

Much more often, regulators will conclude that the costs to privacy outweigh the benefits of unfettered information flow. When they come to such a conclusion, they should consider rules and laws that reduce the risk by restricting the amount of information flowing through society. Of course,

311. For example, Harvard's Personal Genome Project, which is sequencing the DNA of thousands of volunteers to hunt for genetic markers for disease, has essentially told its volunteers to forget about privacy. Peter Dizikes, *Your DNA Is a Snitch*, SALON.COM, Feb. 17, 2009, http://www.salon.com/env/feature/2009/02/17/genetic_testing ("[T]he Personal Genome Project essentially tells its volunteers to forget about privacy guarantees. 'I like the Personal Genome Project approach,' [one scholar] says. 'It's honest. They're saying, "If you want to take the risks, great."'").

such restrictions must be chosen with care because of the important values of free information flow. Regulators should thus try to clamp down on information flow in targeted ways, using the factors listed above in their instrumental sense as a menu of potential interventions.

If the costs significantly outweigh the benefits of information flow, regulators might completely ban the dissemination or storage of a particular type of information. For example, regulators should probably often conclude that public releases of information—even information that seems benign or nonthreatening—should be banned, particularly because such information can be used to supply middle links in long chains of inferences. In more balanced situations, regulators might restrict but not cut off information flow, for example by instituting a quantity cap or a time limit for storage.[312] They might also place even milder restrictions on small classes of trusted people—academic researchers, for example—while banning the sharing of the data with anybody else.

D. Two Case Studies

To demonstrate how a regulator should apply this test, and to highlight the important roles of context and trust, let us revisit again the case studies introduced before: health and internet usage information. Debates about the proper regulation of these two classes of data have raged for many years. Although I cannot capture every nuance of these debates in this space, I revisit them in order to show how to regulate data privacy after the fall of the robust anonymization assumption.

1. Health Information

Once regulators choose to scrap the current HIPAA Privacy Rule—a necessary step given the rule's intrinsic faith in deidentification—how should they protect databases full of sensitive symptoms, diagnoses, and treatments? Consider one class of users of such information in particular: medical researchers seeking new treatments and cures for disease. In this context, both the costs and benefits of unfettered use are enormous. On the one hand, if our worst enemies get hold of our diagnoses and treatments, they can cause us great embarrassment or much worse. On the other hand, researchers use this information to cure disease, ease human suffering, and save lives. Regulators

312. *See* 2008 Working Party Opinion, *supra* note 308, at 19 (arguing search engines should store queries for only six months).

will justifiably be reluctant to throttle information flow too much in this context since the toll of such choices might be measurable in human lives lost.

HIPAA tried to resolve this dilemma by trusting the technology of anonymization. We no longer trust the technology, but we can still rely on a different trust: trust in the researchers themselves. Health researchers are rarely willing to release sensitive data—scrubbed or not—to just anybody who asks. Instead, they tend to share such data only after verifying the bona fides of the person asking. Regulators should build upon such human networks of trust in a revised HIPAA, allowing data transfer where trust is high and forbidding it where trust is low.

The problem is that today researchers trust one another according to informal rules and soft intuitions, and to build trust into law, these rules and intuitions must be formalized and codified. Should HIPAA rely only on a researcher's certification of trust in another, or should an outside body such as an Institutional Review Board review the bases for trust?[313] Should trust in a researcher extend also to her graduate students? To her undergraduate lab assistants? Regulators should work with the medical research community to develop formalized rules for determining and documenting trusted relationships.

Once the rules of verifiable trust are codified, regulators can free up data sharing between trusted parties. To prevent abuse, they should require additional safeguards and accountability mechanisms. For example, they can prescribe new sanctions—possibly even criminal punishment—for those who reidentify. They can also mandate the use of technological mechanisms: both ex ante like encryption and password protection, and ex post review methods like audit trail mechanisms.

Regulators can vary these additional protections according to the sensitivity of the data. For example, for the most sensitive data such as psychotherapy notes and HIV diagnoses, the new HIPAA can mandate an NSA-inspired system of clearances and classifications; HIPAA can require that researchers come to the sensitive data rather than letting the data go to the researchers, requiring physical presence and in-person analysis at the site where the data is hosted. At the other extreme, for databases that contain very little information about patients, perhaps regulators can relax some or all of the additional protections.

While these new, burdensome requirements on their own might stifle research, they would permit another change from the status quo that might instead greatly *expand* research: With the new HIPAA, regulators should

313. According to federal rules, federally-funded research involving human subjects must be approved by an IRB. 45 C.F.R. §§ 46.101–109 (2009).

rescind the current, broken deidentification rules. Researchers who share data according to the new trust-based guidelines will be permitted to share *all* data, even fields of data like birth date or full ZIP code that they cannot access today.[314] With more data and more specific data, researchers will be able to produce more accurate results, and thereby hopefully come to quicker and better conclusions.[315]

This then should be the new HIPAA: Researchers should be allowed to release full, unscrubbed databases to verifiably trusted third parties, subject to new controls on use and new penalties for abuse. Releases to less-trusted third parties should fall, of course, under different rules. For example, trust should not be transitive. Just because Dr. A gives her data to trusted Dr. B does not mean that Dr. B can give the data to Dr. C, who must instead ask Dr. A for the data. Furthermore, releases to nonresearchers such as the marketing arm of a drug company should fall under very different, much more restrictive rules.

2. IP Addresses and Internet Usage Information

Lastly, consider again the debate in the European Union about data containing IP addresses. Recall that every computer on the internet, subject to some important exceptions, possesses a unique IP address that it reveals to every computer with which it communicates. A fierce debate has raged between European privacy advocates who argue that IP addresses should qualify as "personal data" under the Data Protection Directive[316] and online companies, notably Google, who argue that in many cases they should not.[317] European officials have split on the question,[318] with courts and regulators in Sweden[319]

---

314. It makes sense to continue to prohibit the transfer of some data, such as names, home addresses, and photographs that could reveal identity without any outside information at all.

315. The current HIPAA Privacy Rule has itself been blamed for a reduction in data sharing among health researchers.

> In a survey of epidemiologists reported in the Journal of the American Medical Association, two-thirds said the HIPAA Privacy Rule had made research substantially more difficult and added to the costs and uncertainty of their projects. Only one-quarter said the rule had increased privacy and the assurance of confidentiality for patients.

Nancy Ferris, *The Search for John Doe*, GOV'T HEALTH IT, Jan. 26, 2009, http://www.govhealthit.com/Article.aspx?id=71456.

316. 2007 Working Party Opinion, *supra* note 28, at 21; Electronic Privacy Information Center, *Search Engine Privacy*, http://epic.org/privacy/search_engine (last visited Apr. 4, 2010).

317. *See* sources cited *infra* note 324.

318. For a good summary, see Posting of Joseph Cutler, *Was That Your Computer Talking to Me? The EU and IP Addresses as "Personal Data"*, PERKINS COIE DIGESTIBLE LAW BLOG, http://www.perkinscoie.com/ediscovery/blogQ.aspx?entry=5147 (June 24, 2008, 23:30 EST).

and Spain[320] deciding that IP addresses fall within the Directive and those in France,[321] Germany,[322] and the UK[323] finding they do not.

a. Are IP Addresses Personal?

The debate over IP addresses has transcended EU law, as Google has framed its arguments not only in terms of legal compliance but as the best way to balance privacy against ISP need.[324] In this debate, Google has advanced arguments that rely on the now discredited binary idea that typifies the PII mindset: Data can either be identifiable or not. Google argues that data should be considered personal only if it can be tied by the data administrator to one single human being. If instead the data administrator can narrow an IP address down only to a few hundred or even just a few human beings—in other words, even if the administrator can reduce the entropy of the data significantly—Google argues that it should not be regulated. By embracing this idea, Google has downplayed the importance of information entropy, the idea that we can measure and react to imminent privacy violations before they mature.

Google frames this argument in several ways. First, it argues that IP addresses are not personal because they identify machines, not people.[325] Google's Global Privacy Officer, Peter Fleischer, offers hypothetical situations

319. John Oates, *Sweden: IP Addresses are Personal . . . Unless You're a Pirate*, REGISTER, June 18, 2009, *available at* http://www.theregister.co.uk/2009/06/18/sweden_ip_law.

320. AGENCIA ESPAÑOLA DE PROTECCIÓN DE DATOS, STATEMENT ON SEARCH ENGINES (2007), *available at* http://www.samuelparra.com/agpd/canaldocumentacion/recomendaciones/common/pdfs/declaracion_aepd_buscadores_en.pdf (opinion of Spanish Data Protection Agency deciding that search engines process "personal data," relying in part on earlier rulings about IP addresses).

321. Meryem Marzouki, *Is the IP Address Still a Personal Data in France?*, EDRI-GRAM, Sept. 12, 2007, http://www.edri.org/edrigram/number5.17/ip-personal-data-fr.

322. Posting of Jeremy Mittma, *German Court Rules That IP Addresses Are Not Personal Data*, PROSKAUER PRIVACY LAW BLOG, http://privacylaw.proskauer.com/2008/10/articles/european-union/german-court-rules-that-ip-addresses-are-not-personal-data (Oct. 17, 2008).

323. INFO. COMM'R'S OFFICE, DATA PROTECTION GOOD PRACTICE: COLLECTING PERSONAL INFORMATION USING WEBSITES 3 (2007), *available at* http://www.ico.gov.uk/upload/documents/library/data_ protection/practical_application/collecting_personal_information_from_websites_v1.0.pdf.

324. Posting of Alma Whitten, *Are IP Addresses Personal?*, GOOGLE PUBLIC POLICY BLOG, http://googlepublicpolicy.blogspot.com/2008/02/are-ip-addresses-personal.html (Feb. 22, 2008, 12:31 EST) (tying the discussion to the broad question, "as the world's information moves online, how should we protect our privacy?"); Peter Fleischer, *Can a Website Identify a User Based on IP Address?*, PETER FLEISCHER: PRIVACY . . . ?, http://peterfleischer.blogspot.com/2008/02/can-website-identify-user-based-on-ip.html (Feb. 15, 2008) ("Privacy laws should be about protecting identifiable individuals and their information, not about undermining individualization."). Mr. Fleischer serves as Google's Global Privacy Counsel. Because of this, I cite his blog posts for clues about Google's views, but I should be clear that Mr. Fleischer's blog bears the disclaimer, "these ruminations are mine, not Google's."

325. *Cf.* Fleischer, *supra* note 324 (An IP address "constitutes by no means an indirectly nominative data of the person in that it only relates to a machine, and not to the individual who is using the computer in order to commit counterfeit." (quoting decision of the Paris Appeals Court)).

in which many users share one computer with a single IP address, such as "the members of an extended family each making use of a home pc, a whole student body utilising a library computer terminal, or potentially thousands of people purchasing from a networked vending machine."[326] Is Fleischer right to categorically dismiss the threat to privacy in these situations? Is there no threat to privacy when Google knows that specific search queries can be narrowed down to the six, seven, maybe eight members of an extended family? For that matter, should regulators ignore the privacy of data that can be narrowed down to the students on a particular college campus, as Fleischer implies they should?

Second, in addition to the machine-not-person argument, Google further ignores the lessons of easy reidentification by assuming it has no access to information that it can use to tie IP addresses to identity. On Google's official policy blog, Software Engineer Alma Whitten, a well-regarded computer scientist, asserts that "IP addresses recorded by every website on the planet *without additional information* should not be considered personal data, because these websites usually cannot identify the human beings behind these number strings."[327] Whitten's argument ignores the fact that the world is awash in rich outside information helpful for tying IP addresses to places and individuals.

For example, websites like Google never store IP addresses devoid of context; instead, they store them connected to identity or behavior. Google probably knows from its log files, for example, that an IP address was used to access a particular email or calendar account, edit a particular word processing document, or send particular search queries to its search engine. By analyzing the connections woven throughout this mass of information, Google can draw some very accurate conclusions about the person linked to any particular IP address.[328]

Other parties can often link IP addresses to identity as well. Cable and telephone companies maintain databases that associate IP addresses directly to names, addresses, and credit card numbers.[329] That Google does not store these data associations on its own servers is hardly the point. Otherwise, national

---

326. Peter Fleischer, *Are IP Addresses "Personal Data"?*, PETER FLEISCHER: PRIVACY…?, http://peterfleischer.blogspot.com/2007/02/are-ip-addresses-personal-data.html (Feb. 5, 2007, 17:18 EST).

327. Whitten, *supra* note 324 (emphasis added).

328. *See* 2008 Working Party Opinion, *supra* note 308, at 21 ("The correlation of customer behaviour across different personalised services of a search engine provider . . . can also be accomplished by other means, based on . . . other distinguishing characteristics, such as individual IP addresses.").

329. *Id.* at 11, 16.

ID numbers in the hands of private parties would not be "personal data" because only the government can authoritatively map these numbers to identities.[330]

Google can find entropy-reducing information that narrows IP addresses to identity in many other places: Public databases reveal which ISP owns an IP address[331] and sometimes even narrow down an address to a geographic region;[332] IT departments often post detailed network diagrams linking IP addresses to individual offices; and geolocation services try to isolate IP addresses to a particular spot on the Earth.[333] In light of the richness of outside information relating to IP addresses, and given the power of reidentification, Google's arguments amount to overstatements and legalistic evasions.

Google's argument that it protects privacy further by deleting a single octet of information from IP addresses is even more disappointingly facile and incorrect. An adversary who is missing only one of an IP address's four octets can narrow the world down to only 256 possible IP addresses.[334] Google deserves no credit whatsoever for deleting partial IP addresses; if there is a risk to storing IP addresses at all, Google has done almost nothing to reduce that risk, and regulators should ask them at the very least to discard all IP addresses associated with search queries, following the practice of their search-engine competitors, Microsoft and Yahoo.[335]

b. Should the Data Protection Directive Cover Search Queries?

Not only does the easy reidentification result highlight the flaws in Google's argument that IP addresses are not personal, it also suggests that European courts should rule that the EU Directive covers IP addresses. Recall that the Directive applies broadly to any data in which a "person . . . can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological,

330. Fleischer correctly points out that ISPs are often forbidden from disclosing the user associated with an IP address. Fleischer, *supra* note 324 ("[T]he ISP is prohibited under US law from giving Google that information, and there are similar legal prohibitions under European laws.") This is no different from any other kind of account number which can be authoritatively tied to identity only by the issuing entity. All other entities must make educated guesses.

331. *E.g.*, ARIN WHOIS Database Search, http://ws.arin.net/whois (last visited June 12, 2010) ("ARIN's WHOIS service provides a mechanism for finding contact and registration information for resources registered with ARIN.").

332. ERIC COLE & RONALD KRUTZ, NETWORK SECURITY BIBLE 316–18 (2005) (discussing reverse DNS queries).

333. *E.g.*, IP2Location.com, http://www.ip2location.com (last visited June 12, 2010); Quova, http://www.quova.com (last visited June 12, 2010).

334. An octet is so named because it contains eight bits of data. $2^8 = 256$.

335. *See supra* note 200.

mental, economic, cultural or social identity."[336] Because websites can often tie IP addresses to individual people, the Directive should apply to them. Still, courts in Germany, France, and the UK have held to the contrary. Should the EU amend the Directive to even more unequivocally cover IP addresses?

The answer is not to expand the Directive to specifically cover IP addresses, as we might have done when we still organized laws solely around PII. Instead, the EU should enact new, sectoral regulations that reflect a weighing of costs and benefits for specific problems. In this case, rather than ask whether any company holding an IP address should bear the burden of the EU Directive, the EU might ask whether the benefit of allowing search engines in particular to store and disclose information—including IP addresses associated with search queries—outweighs the potential harm to privacy.[337]

I must save for another day a complete response to this question, but to demonstrate the new test for deciding when to regulate after the fall of anonymization, I will outline why I think search engines deserve to be regulated closely. Compare the benefits and costs of allowing unfettered transfers of stored search queries to the earlier discussion about health information, taking the benefits first. By analyzing search queries, researchers and companies can improve and protect services, increase access to information, and tailor online experiences better to personal behavior and preferences.[338] These are important benefits, but not nearly as important as improving health and saving human lives.

On the other side of the ledger, the costs to privacy of unfettered access are probably as great for search query information as for health information, if not greater. As the AOL breach revealed, stored search queries often contain user-reported health symptoms.[339] In fact, Google takes advantage of this to track and map influenza outbreaks in the U.S.[340] When one considers how often Google users tell Google about symptoms that never escalate to a visit to the doctor, one can see how much richer—and thus more sensitive—this information can be than even hospital data.

We reveal even more than health information to search engines, supplying them with our sensitive thoughts, ideas, and behavior, mixed in of course with

336. EU Data Protection Directive, *supra* note 3, art. I(a).

337. In the EU, the Article 29 Working Group privacy watchdog has proposed similarly special treatment for search engines. 2008 Working Party Opinion, *supra* note 308, at 24.

338. *See supra* note 201.

339. Barbaro & Zeller, *supra* note 69 ("Her search history includes 'hand tremors,' 'nicotine effects on the body,' 'dry mouth' and 'bipolar.' But in an interview, Ms. Arnold said she routinely researched medical conditions for her friends to assuage their anxieties.").

340. Google.org, Flu Trends, http://www.google.org/flutrends (last visited June 12, 2010).

torrents of the mundane and unthreatening.[341] In an earlier article, I argued that the scrutiny of internet usage—in that case by Internet Service Providers—represents the single greatest threat to privacy in society today.[342] Regulators have underappreciated the sensitive nature of this data, but events like the AOL data release have reawakened them to the special quality of stored search queries.[343]

Because the costs of unfettered data access are as high in the search-engine as in the health context, EU and U.S. regulators should consider enacting specific laws to govern the storage and transfer of this information. Because the benefits are less than for health information, regulators should be willing to restrict the storage and flow of search query information even more than HIPAA restricts health information.

Thus, the EU and U.S. should enact new internet privacy laws that focus on both the storage and transfer of search queries. They should impose a quantity cap, mandating that companies store search queries for no longer than a prescribed time.[344] They should set the specific time limit after considering search companies' claims that they must keep at least a few months' worth of data to serve vital business needs. They should also significantly limit third-party access to search query data.

CONCLUSION

Easy reidentification represents a sea change not only in technology but in our understanding of privacy. It undermines decades of assumptions about robust anonymization, assumptions that have charted the course for business relationships, individual choices, and government regulations. Regulators must respond rapidly and forcefully to this disruptive technological shift, to restore balance to the law and protect all of us from imminent, significant harm. They must do this without leaning on the easy-to-apply, appealingly nondisruptive, but hopelessly flawed crutch of personally identifiable information. This Article offers the difficult but necessary way forward: Regulators must use the factors provided to assess the risks of reidentification and carefully balance these risks against countervailing values.

---

341. *Cf.* Julie Cohen, *Examined Lives: Informational Privacy and the Subject as Object*, 52 STAN. L. REV. 1373, 1426 (2000).

342. Paul Ohm, *The Rise and Fall of ISP Surveillance*, 2009 U. ILL. L. REV. 1417, 1417.

343. *See* 2008 Working Party Opinion, *supra* note 308, at 8 ("Search engines play a crucial role as a first point of contact to access information freely on the internet.").

344. *Cf. id.* at 19 ("[T]he Working Party does not see a basis for a retention period beyond 6 months.").

Although reidentification science poses significant new challenges, it also lifts the veil that for too long has obscured privacy debates. By focusing regulators and other participants in these debates much more sharply on the costs and benefits of unfettered information flow, reidentification will make us answer questions we have too long avoided. We face new challenges, indeed, but we should embrace this opportunity to reexamine old privacy questions under a powerful new light.